Examining the Impact of Examinee-Selected Constructed Response Items in the Context of a Hierarchical Rater Signal Detection Model

Brian F. Patterson

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

ABSTRACT

Examining the Impact of Examinee-Selected
Constructed Response Items in the Context of a
Hierarchical Rater Signal Detection Model

Brian F. Patterson

Research into the relatively rarely used examinee-selected item assessment designs has revealed certain challenges. This study aims to more comprehensively re-examine the key issues around examinee-selected items under a modern model for constructed-response scoring. Specifically, data were simulated under the hierarchical rater model with signal detection theory rater components (HRM-SDT; DeCarlo, Kim, & Johnson, 2011) and a variety of examinee-item selection mechanisms were considered. These conditions varied from the hypothetical baseline condition—where examinees choose randomly and with equal frequency from a pair of item prompts—to the perhaps more realistic and certainly more troublesome condition where examinees select items based on the very subject-area proficiency that the instrument intends to measure. While good examinee, item, and rater parameter recovery was apparent in the former condition for the HRM-SDT, serious issues with item and rater parameter estimation were apparent in the latter. Additional conditions were considered, as well as competing psychometric models for the estimation of examinee proficiency. Finally, practical implications of using examinee-selected item designs are given, as well as future directions for research.

CONTENTS

TABLES

FIGURES

Figure **Page**

ACKNOWLEDGEMENTS

I gratefully acknowledge the immense contribution that my advisor Dr. Lawrence T. DeCarlo has made, not only to this current research endeavor, but to the larger accumulation of professional skills and knowledge that I have made in my time as his student. He spent countless hours reviewing manuscript drafts and counseling our research group on the best and most effective ways to conduct and present research. For this I cannot thank him enough. His clear and passionate regard for his students is a testament to his desire to shape the next generation of researchers and practitioners toward making more fruitful contributions to academia and industry. Also, I would be remiss in not thanking the rest of the faculty and staff at Teachers' College, especially my dissertation committee members: Drs. James Corter, Young-Sun Lee, and Steven Peverly, and my Columbia University examiner Dr. Liam Paninski.

The next layer of support that I received was from my wife Susan, our son Colin, my brother Jeffrey, my parents Jill and Tom, and my family by marriage, Don, Theresa, and Kathy Sullivan. They always supported me, were understanding when I could not devote to them the time that they deserved, and made the best of what seemingly little time we did have together. It was the knowledge of their earnest support that kept me motivated, especially in the early stages of this work when I struggled to formulate my research questions and planned approach.

Another group that was critical to my success was my College Board colleagues. My supervisor Dr. Mary-Margaret Kerns gave me the flexibility and encouragement needed to complete this doctoral program while working full-time. And it was my supervisor at the time I was hired, Dr. Kristen Huff, who encouraged me to pursue my study of measurement and statistics. Dr. Huff, along with so many others—including Drs. Wayne Camara, Michael Chajewski, and F. Tony Di Giacomo—always had an encouraging word or humorous anecdote about their graduate studies. I humbly thank these wonderful people and am honored to count them all as friends.

DEDICATION

To my incredibly understanding wife Susan, without whom this would not have been possible.

Chapter I

INTRODUCTION

In the vast majority of educational assessments, examinees are expected to respond to all items. There are, however, a small number of situations in which a test publisher allows examinees individually to respond to only a subset of items. For example, examinees may be presented with five possible essay topics and instructed to respond to any three. Willmott and Hall (1975) traced the use of such examinee-selected items as far back as the 1858 Oxford Associated Arts Examination; they quoted the Oxford Delegacy of Local Examinations' records of a report on that assessment:

> "It may be remarked generally, that a much larger number of questions was set than any one candidate was expected to answer, and that questions suited to the younger and older candidates were included in the same paper. This arrangement was adopted in order to give every candidate the widest range of selection, and the least occasion for subsequent complaint. Considering the various circumstances of the youths, no other plan presented so little prospect of inconvenience, and no inconvenience did in fact result from the course adopted." (Willmott & Hall, 1975, p. 6)

While it may be argued that some inconvenience may have been present, but undetected, the use of examinee-selected items persists. Most often, examinee-selected item designs are employed in the context of longer essays or problems, as opposed to shorter item formats like multiple choice or true-false items. These essays, lengthy science problems, and other items where the examinee does not simply select from a set of response options are referred to as constructed response (CR) items, though they are also known as subjective or examinee-produced response items.

1.1. Examples of and Settings for Using Examinee-Selected Items

Examples of exams that use examinee-selected CR items exist in a variety of testing programs. The Advanced Placement® (AP®) exams in U.S. and European history (College Board, 2010a, 2010b) and earlier forms of the AP chemistry exam (College Board, 2006) utilized examinee-selected CR items. The Advanced Placement history exam offered examinees a different choice design from those in chemistry. While AP examinees in history choose one each from two different sets of items (i.e., testlets), an earlier version of the AP exam in chemistry instructed examinees to select three out of a pool of five possible items to which to respond (Lukhele, Thissen, & Wainer, 1994). Additional examples of testing programs that at some point have employed examinee-selected items include the United Kingdom's General Certificate of Secondary Education (GCSE) exam in English literature (OCR, 2003); the Maryland School Performance Assessment Program, Grades 3, 5, and 8 (Fitzpatrick & Yen, 1995); and the Michigan English Language Assessment Battery (MELAB; Hamp-Lyons & Mathias, 1994). Other testing programs have studied the use of examinee-selected items, either in experimental settings or for small subsets of operational administrations.

Examinee selection of test items may be appropriate in certain settings in norm-referenced assessments, especially when the test questions primarily serve as stimuli for examinees to demonstrate general skills, such as structuring an argument. In other words, in such settings, the deeper construct may be measured equally well by items with varying surface content. Consider, for example, a situation examined by Campbell and Donahue (1997) in which a literature instructor taught a core set of skills through the analysis of a variety of different texts. In such a case, a test publisher may allow for examinees to choose the reading passage on which they must write essays. In this situation, the test would assess mastery of targeted analytical skills and essay-writing proficiency regardless of which reading passage a student chose.

In the history domain, teachers may place differential emphasis on certain historical periods or events, depending upon their knowledge and their students' interests. The instructors might use these different foci to cultivate the same general historical skills in their students. An assessment could reasonably allow for examinee-selection of items that enable examinees to display their historical analytic skills independent of the particular context in which they are examined. Here again, the surface content of the item is less important than the measurement of the deeper construct.

Examinee-selected items may not be appropriate, however, in cases where typical performance on a representative set of tasks, for example in professional licensure testing. In such a situation, an accrediting body expects professionals to be able to carry out certain job functions that are required of successful professionals in that field. For example, all candidates for medical licensing may be expected to be able to successfully apply surgical sutures and draw blood samples, so the accrediting body probably should not offer candidates the choice to either apply sutures or draw blood samples. Since professionals should not expect to be able to exert control over the representative job tasks that they must undertake, it may not be reasonable to grant them such control over the content of the certifying, criterion-referenced test.

1.2. Possible Consequences of Using Examinee-Selected Items

The use of examinee-selected items has a variety of expected benefits, but there are also some well-reasoned arguments against their use. In surveying the existing literature on examinee-selected items, a number of theoretical claims recur throughout. Those in favor of examinee-selected items note that their use (a) enables examinees to maximize their expected scores (Bridgeman, Morgan, & Wang, 1997); (b) enhances fairness (Allen, Holland, & Thayer, 2005; Wainer & Thissen, 1994); and (c) frees the classroom teacher from focusing too much on

the summative or end-of-course exam (Bell, 1997; Wainer & Thissen, 1994). Of course there may be problems with the use of examinee-selected items, chief among them are (a) the possibly incommensurable scores among examinees choosing different items (Wainer et al., 1994; Wang, Wainer, & Thissen, 1995); and (b) the real possibility that examinees may not choose the item that would maximize their scores (Bridgeman et al., 1997).

In addition to these theoretical claims and empirical studies, it is important to consider the unique psychological processes that take place in the examinees' minds when presented with examinee-selected items. Rational examinees, when presented with a choice of CR items, should choose the item that will maximize their final scaled score. All other things being equal, they should choose the item that (a) is easiest; and/or (b) will be graded most leniently. This is clearly a situation of decision under uncertainty, as examinees may not be able to estimate (a) their own proficiency for the tasks before them; (b) the true item difficulty; or (c) the stringency with which raters will grade their work. Given this uncertainty, a natural concern is that, as Kruger and Dunning (1999) showed, individuals may overestimate their own proficiency in the area being tested and therefore may select an item that could disadvantage them. For example, they may choose the harder item in the mistaken belief that they have developed sufficient skill and knowledge to answer adequately, when in fact they could have scored better on another item. Above and beyond their knowledge of their own proficiency, they would be hard-pressed to be able to precisely detect differences in item difficulty and it would be virtually impossible for them to predict which item would be graded less stringently. These issues around the examinees' choice itself cannot be ignored when considering this practice.

1.3. Aim of the Current Study

The psychometric case for examinee-selected items seems at best inconclusive or at worst non-existent. The above just scratches the surface of arguments and evidence for and

against the use of examinee-selected items. Such argumentation and empirical results have not adequately illustrated the consequences of using examinee-selected items for the estimation of psychometric models, the prediction of examinee proficiency, or the characterization of rater (i.e., the human judge of CR quality) traits. In other words, while clear evidence and sound reasoning have been presented for some of these points, no comprehensive evaluation of hypothetical examinee item selection methods has been performed under a modern constructed response modeling framework.

In addition, there are relatively few studies (with the exception of Bradlow & Thomas, 1998) that put forth a theoretical basis for how examinees select items. There are any number of possible decision-making strategies that examinees might use when presented with examinee-selected items. This study will simulate a handful of likely situations, selected to cover a spectrum of hypothetical selection mechanisms. Particular emphasis will be given to the situations in which the selection mechanism may have the most severe impact on the estimation of the chosen psychometric model. In particular, examinee traits will be generated, hypothetical assessment items will "administered" and "rated," and three possible selection mechanisms will be simulated. In one condition, examinees' item selection will be unrelated to the proficiency dimension (i.e., the examinee's "ability") that the assessment aims to measure. In another, so-called "test-wise" individuals will select items differently from those lacking test wisdom, but who may be otherwise similar in terms of underlying proficiency. In the final condition, examinees above a certain proficiency threshold will choose the easier item, with less proficient examinees unable to distinguish between the items' difficulties. The model for constructed response scores will be estimated in each of these conditions and for a number of independently simulated datasets and estimates will be compared with known parameter values.

One useful framework for the estimation of examinee, rater, and item characteristics is the hierarchical rater model with signal detection theory rater components (HRM-SDT; DeCarlo,

Kim, & Johnson, 2011). This model has a variety of advantages over competing models, including that it enables the separate estimation of item and rater parameters. The rater model is grounded in signal detection theory (SDT; Wickens, 2002) in order to represent the psychological process of rating CRs. Differences in rater discrimination (i.e., detection) and rating stringency; item discrimination and thresholds; and examinee proficiency are all represented by distinct parameters. It also appropriately characterizes individual raters' judgments as indirect, rather than direct, indicators of examinees' proficiency (DeCarlo et al., 2011). For these reasons, the HRM-SDT will be used as the population model for data generation and it will be the main focus in the consideration of the effects of examinee-selected items.

It is through the lens of the HRM-SDT that the effects of examinee item selection will be appraised critically and new light will be shed on questions that were not answered cohesively in prior studies. The following are the primary questions to be answered:

1. Does the use of examinee-selected items lead to predictions of examinee proficiency with differential precision (i.e., different variance) or differential accuracy (i.e., different bias)? Under which examinee selection mechanisms are these issues most prominent?
2. Can the HRM-SDT precisely (i.e., with reasonable variance) and accurately (i.e., with little bias) estimate item and rater characteristics, even in the presence of examinee-selected items? Again, does this vary across different examinee item selection mechanisms?

Examinee-selected items are currently used to ensure fairness, but since the late 1990s relatively little effort has been expended to evaluate newly developed models' performance or develop novel approaches that might mitigate the problems associated with selection. It is with this in mind and a comprehensive constructed response modeling framework in hand that this study will evaluate the feasibility of using examinee-selected constructed response items.

Chapter II

LITERATURE REVIEW

2.1. Examinee-Selected Items

As Bridgeman et al. (1997) insightfully noted, "…given limited testing time and scoring resources, someone must choose the essay topic; the question is whether the test designer or examinee should get to make the choice" (p. 283). Clearly the authors did not intend for their point to be taken to the extreme—where examinees' construct an entire assessment from an infinite menu of possible items—however, to the extent that allowing examinee choice is psychometrically feasible and appropriate for what is being measured, it is worth considering. Examinee-selected items are most often implemented in the context of constructed response (CR) items. The majority of the studies to follow (summarized in Table 1) examine examinee-selected CR items, rather than objective response formats, like multiple choice (MC) items.

Table 1.

Summary of Benefits and Limitations of Examinee-Selected Items

Benefit	Claim / Evidence	Reference
1. Fairness	▪ This is an anecdotal claim with no apparent operationalization.	Allen, Holland, & Thayer (2005); Wainer & Thissen (1994)
	▪ Subgroup differences may be reduced, for example the Black-White score gap.	Gabrielson, Gordon, & Engelhard Jr. (1995)
2. Content Validity Balanced with Practical Concerns	▪ Classroom teachers are freer to cover topics that they understand best and/or most interest students.	Wainer & Thissen (1994); Bell (1997)
	▪ Course may be broad enough to cover key content and skills, while not requiring overly long exams.	Wainer & Thissen (1994)
3. Examinees Maximize Ratings	▪ Assumes rational, self-aware examinees.	Bridgeman, Morgan, & Wang (1997)
	▪ In a choose-one, answer-all design, most examinees chose the topic that maximized their essay rating.	
	▪ Examinees choosing a topic tend to out-perform those assigned to it.	Allen, Holland, & Thayer (2005); Jennings, Fox, Graves, & Shohamy (1999)

Challenge	Claim / Evidence	Reference
1. Psychometric Challenges	▪ Scores based on examinee-selected items may not comparable or unidimensional.	Wainer & Thissen (1994)
	▪ Items may be differentially difficult	
	▪ If item difficulty varies greatly, item and examinee parameter estimates may be biased.	Wainer & Thissen (1994); Bradlow & Thomas (1998)
2. Examinees May Select Poorly	▪ Found no significant differences in essay ratings by topic chosen.	Bridgeman, Morgan, & Wang (1997); Gabrielson, Gordon, & Engelhard Jr. (1995)
	▪ Some examinees do not select the item that would have maximized their score.	Gabrielson, Gordon, & Engelhard Jr. (1995)
3. Proficiency May Drive Item Selection	▪ Examinees with higher proficiency are more likely to choose the item to maximize their CR ratings.	Fitzpatrick & Yen (1995); Sarnacki (1979); Chi (1978)

Expected Benefits of Examinee-Selected Items. As summarized in Table 1, research into examinee-selected assessment items involves conflicting claims about several purported benefits. Examinees may have divergent mastery of the topics to be examined; a large part of that disparity may be caused by the varied experience and interests of their classroom teachers. In those classrooms where teachers focus on one historical period at the expense of others, for example, examinees are constrained by the choices of their teachers, well-intentioned though they may be. As such, when the class is nearing completion and the summative exam is approaching, giving students a choice in the essay topics on which they may write is presumed to be a fair method of ensuring that the examinees are able to demonstrate their mastery of general analytic skills in the historical domain. Thus, it is anecdotally observed that the use of examinee-selected items is motivated primarily by claims of "fairness" (Allen, Holland, & Thayer, 2005; Wainer & Thissen, 1994).

In addition to providing benefits for students, Wainer and Thissen (1994) make the compelling point that enabling examinees to choose from among a set of essay topics frees the classroom teacher somewhat to focus on the areas in which he or she has particular expertise and interest. The use of examinee-selected responses also frees the test publisher from having to design an exam that both measures all content covered in an entire academic year and remains a reasonable length. Examinee-selected items are considered a practical solution to keep test length feasible without narrowing the scope of a course to what may be examined in a single test administration (Wainer & Thissen, 1994). In other words, the use of examinee-selected items may balance the desire for content validity (i.e., greater coverage of course topics) against more practical limitations (e.g., limited instructional time, fixed exam period). Such intellectual freedom may have the additional advantage of better motivating both teachers and students and may free them from strictly focusing on what is expected to be examined at the conclusion of the course. The use of examinee-selected items may thus ensure that

measurement is driven by instruction (Bell, 1997), rather than measurement driving instruction (i.e., “teaching to the test”).

Despite the best efforts of test publishers to ensure uniform coverage of content, differences in teacher experience may lead to variations in the coverage of particular historical, political, or economic topics relevant to AP United States (U.S.) and European history exams. Therefore it is possible that students may be differentially well prepared to respond to certain topics. Examinees who are fortunate enough to be asked on the summative exam questions about content that their instructors have covered well may be perceived to have an unfair advantage.

In addition to being seen by many as a fair practice, there is also the expectation that examinees will perform better on average when presented with choices, as opposed to being instructed to answer all items. Bridgeman et al. (1997) conducted a study analyzing examinee-selected items in what they called a “choose-one, answer all” (C1-A) design. In that study, the authors presented participants with two of a possible four essay topics at which point examinees “were told that they should first choose their preferred topic, although they should answer both and both would be scored” (Bridgeman et al., 1997, p. 275). In particular, the authors instructed examinees to read both essay topics and select the one that they were best prepared to answer thoroughly. In doing so, the authors were able to determine whether the examinees’ choices led to maximizing their exam scores. The authors showed that in most cases—between 61% and 92%, depending upon preferred topic—participants did choose the topic on which they would achieve the highest ratings.

In a similar vein, Jennings, Fox, Graves, and Shohamy (1999) conducted an experiment comparing examinees who either were or were not given a choice of essays in the context of the Canadian Academic English Language (CAEL) Assessment. Since CAEL is a topic-based assessment—meaning that the reading, listening (i.e., lecture), and writing portions all relate to

a single topic—the authors wanted to rule out the possibility that prior knowledge of the topic had an impact on examinees' demonstrated mastery of English for use in an academic setting. Jennings et al. (1999) found that, for the same topic, the examinees who could choose a topic tended to earn slightly higher English proficiency ratings than those who were assigned to the same thematic topic. It is unsurprising that there was only a small effect, since English communication was the object of assessment (i.e., the construct being measured), and it was only the surface item content that varied.

Powers and Bennett (1999) also found that examinees performed better when choosing an item than when assigned to it. They used an experimental section added to the Graduate Record Examinations (GRE), either assigning examinees to all three of six prompts or assigning them to two and allowing the choice of one from three possible prompts. Those choosing the prompt showed small (i.e., 0.1–0.2 *SD*) but significantly ($p < .001$) higher scores on the prompt than those assigned to it (p. 271-273).

Not only has the use of examinee-selected items shown promise in enabling examinees to maximize essay ratings, but it has been associated with reduced gender; racial / ethnic identity; or other performance gaps. One possible explanation for possible decreases in sub-group performance gaps could be that the use of examinee-selected items could increase achievement motivation, reducing any cross-group motivation gaps. By appealing to a wide variety of possible topics and thereby touching on examinees' academic interests (Fitzpatrick & Yen, 1995), it is expected that examinees will be motivated to produce their best work.

The impact of examinee-selected CR items was examined for gender and racial / ethnic identity by Gabrielson, Gordon, and Engelhard Jr. (1995). High school students taking Georgia's state assessment under an examinee-selected condition exhibited a similar gender difference—with women out-performing men—to those under an assigned topic condition, in terms of ratings on a persuasive writing task. On the other hand, the Black-White score gap was slightly smaller

when examinees were given a choice rather than being assigned to a topic—with standardized differences (i.e., effect sizes) ranging from 0.04 to 0.08 smaller in the choice condition, rather than the assigned topic condition (p. 282).

There has been some evidence put forth indicating that the use of examinee-selected items may improve important psychometric properties of an exam, in particular, validity. There are some who argue that the use of examinee-selected items actually adds construct-irrelevant variance (Messick, 1989), but it seems more reasonable to suspect the opposite. The line of reasoning that Linn, Betebenner, and Wheeler (1998) propose is that the use of examinee-selected items reduces construct-irrelevant variance by enabling examinees to select items that more closely align with their classroom experiences. If the intended use of test scores were, for example, to determine the level of mastery that a candidate exhibits in general historical skills and abilities, then designs like that used in the Advanced Placement U.S. and European history exams would likely lead to more valid inferences. Consider for example, the examinee who is extremely proficient in terms of general historical skills, but simply was not exposed to a given topic area that was required on a summative assessment. The requirement to address an area outside the scope of the examinee's study introduces construct irrelevant variance, if indeed the construct is general historical skills.

Finally, some studies (e.g., Fitzpatrick & Yen, 1995; Powers & Bennett, 1999) found that correlations of examinee-selected items with the total test score—a relationship that would be expected to be strong for a reliable measure—were markedly higher than those of required items with total test score. In other words, the choice items were more internally consistent with the total test (i.e., more reliable) than were the required items. This finding that examinee-selected items may be lead to more reliable test scores is counter-intuitive, since it potentially makes test scores less comparable across examinees selecting different items.

Expected Limitations of Examinee-Selected Items. Despite the apparent benefits of using examinee-selected items, there are certainly possible challenges. With proponents commonly arguing that the use of examinee-selected items allows candidates to maximize their performance, empirical evidence is needed. Gabrielson et al. (1995) used a multivariate analysis of variance model and found no significant difference in essay scores by topic chosen. The authors showed in a study of persuasive writing in a high-stakes setting that, under a choice of two possible topics, examinees tended to produce essays that scored no better than those who were assigned an essay topic. Indeed there are situations where examinees made the wrong choice, performing more poorly than those who were assigned to a task (Gabrielson et al., 1995). It should be noted that this may be attributable to the fact that (a) all tasks were equally difficult and hence choice did not enable examinees to maximize their expected score; (b) examinees were unable to identify the item that would yield the highest essay rating; or (c) examinees were harmed (e.g., made more anxious or required to spend time choosing) by the presentation of the novelty of a choice on an exam (Gabrielson et al., 1995).

In the context of the English Literature qualifying GCSE exam, Bell (1997) estimated item, examinee, and test-center effects. Using a three-level hierarchical linear model, selected essays were nested within examinees, which in turn were nested within testing centers. It is important to note that the test center staff assigned the literary work—exercising some degree of choice—and that the examinee then selected from items within that work. Bell observed that the order in which essay prompts were presented appeared to be negatively related to the proportion of examinees choosing each item (i.e., there seemed to be a tendency to choose the earlier items). He also found small question effects, after controlling for testing center and examinee random intercept effects. However, this study gives little additional evidence for or against the use of examinee-selected items because it does not acknowledge the possible relationship between selected item and latent proficiency. In other words, if as Chi (1978)

suggests, higher proficiency examinees can better detect item difficulty, then Bell's (1997) results may be called into question.

Differential difficulty may exist among items from which examinees select and therefore may increase the stakes associated with choosing the “best” item. Bridgeman, Morgan, and Wang (1997), in whose choose-one, answer-all experimental setting examinees were randomly assigned to spiraled pairs of four possible items, observed differences in mean CR item scores across topics. Allen et al. (2005), who re-analyzed data from Bridgeman et al.(1997), concluded that those mean differences must be attributable to either (a) differential content coverage (i.e., course-item alignment); or (b) different rater grading standards, across items. While they conclude that the items do differ on mean raw score, they are notably silent on the very real possibility that the four items considered across conditions may in fact vary on their innate difficulty, independent of rater idiosyncrasies. In other words, the authors could not rule out differences in item difficulty as causing the differences in observed scores, so any claims about differences in examinee proficiency are limited.

Beyond differences in item difficulty, another possible concern is the structure of the latent traits that tapped by the examinee-selected items. In particular, Wainer and Thissen (1994) propose that the argument for examinee-selected items requires the positing of multiple dimensions of proficiency and that if examinee-selected items tap different dimensions, they would yield non-comparable scores. Consider for example, a pair of examinee-selected U.S. history items: one addressing the civil war and another addressing the civil rights movement. Wainer and Thissen (1994) argue that if the test publisher’s argument for using examinee-selected items is that examinees may possess differential knowledge on the two topics then the test publisher is implicitly assuming the presence of multiple dimensions of proficiency. While this point cannot be denied, with good specification of target item difficulty and relative content coverage (i.e., “test blueprints”) and a common scoring rubric for all prompts in a given choice

set (Bell, 1997), test publishers ought to be able to produce essentially unidimensional examinee-selected item sets. That is, while the surface content of items may vary within examinee-selected item sets, the same deeper construct should be measured by all items within the set subject to examinee selection. In this way, test publishers have the best chance of producing tests that primarily measure a single deep construct.

Setting aside multidimensionality for a moment, suppose that examinees generally do maximize their CR scores when presented with examinee-selected items as proponents suggest they do. That requires that either (a) items differ greatly in terms of difficulty; or (b) raters differ substantially in terms of their leniency. Little evidence and few arguments have been presented to the differential rater effects, but some authors have discussed equating scores to address differential item difficulty and thereby make score scales more comparable (Wainer & Thissen, 1994). The argument for equating, given unidimensionality and with the primary goal of having examinees maximize their scores, is hard to make. Assuming minimal rater effects, if the test is unidimensional and examinees would perform substantially differently, then the items must be differentially difficult. Equating would essentially adjust for that differential difficulty and thereby remove the need for choice (Wainer & Thissen, 1994). Another possibility is that the test publisher may consider the examinee item selection itself to be a component of the construct being examined. Perhaps this is less of an issue, as Wang, Wainer, and Thissen (1995) note, “if one considers the test to be all of the items and the choice, then everyone has received the same form and no equating is necessary” (p. 212).

This notion that items and the selection of items represent a single construct is not without issues. It is a line of reasoning that leads to the expectation that examinees’ standing on the latent proficiency construct may drive the choice of items. However, this assertion is not universally true. For example, Fitzpatrick and Yen (1995) found that in grades 3 and 5 examinees who chose the items whose estimated IRT parameters indicated that they were the

easiest tended to earn below-average test scores. In other words, lower-proficiency examinees chose to respond to the easier item. The authors did not find such a clear pattern for their grade 8 sample; in this sample, lower-proficiency examinees tended to choose one of the easier items, but higher-proficiency examinees also chose an easy item.

Just as examinee proficiency and item selection represent different, but related traits, so too are item and rater characteristics important and separate aspects to be modeled. Existing research on examinee-selected items have examined the effects of that practice on either observed scores or predicted proficiency from an IRT model perspective. Few have attempted to separate rater and item effects, but Gee (1987) found that on average, examinees earned higher scores when they selected the less frequently chosen topics. The author attributed this to raters being more severe on the most commonly selected topics as a result of their frustration from rating so many essays on the same topic (Gee, 1987, p. 103).

To this point, all studies of examinee-selected items discussed in this chapter have focused on CR items, but there was one seminal study that examined the consequences of using examinee-selected multiple choice (MC) items. Aside from rater effects, many of the above issues are expected to carry over into the MC item context, and it is reasonable to expect that key findings from the MC item format may apply for CR items. Bradlow and Thomas (1998) simulated a complete dataset and deleted simulated item scores (i.e., introduced missingness) to represent three possible ways in which examinees choose among pairs of multiple choice items. In one condition—called missing completely at random (MCAR)—examinees' response to one of each pair of items was randomly set to missing. This would correspond to examinees arbitrarily choosing which item to respond to, independent of their own proficiency and of the item characteristics. By definition, the missingness being unrelated to examinee proficiency, it was as if examinees were randomly assigned to answer items. More relevant to this study was the authors' missing not at random (MNAR) condition where, with probability .95, examinees

with proficiency (i.e., θ) greater than zero selected the easier item and with probability .95 examinees with θ < 0 selected the harder item. In other words, examinees with higher proficiency were expected to answer the easier item 95% of the time and those with lower proficiency were expected to answer the harder item 95% of the time.

Bradlow and Thomas (1998) demonstrated that under the MCAR condition, there was negligible bias in the recovered item difficulty parameters, but in the MNAR condition where item choice depended upon θ, the bias of the difficulty parameter estimates increased as a function of the known difficulty values. In other words, there was negative bias for known difficulty values that were large and negative (i.e., substantially underestimated) and positive bias for difficulty parameter estimates whose known values were large and positive (i.e., substantially overestimated). Recovery of examinee proficiency under those two conditions followed the same general pattern, with no bias in the MCAR condition and substantial bias increasing with (i.e., positively related to) known proficiency.

Bradlow and Thomas (1998) concluded that standard unidimensional IRT methods may fail to yield unbiased estimates of item and examinee parameters under examinee-selected items if examinee proficiency affects item choice in a manner similar to what was simulated in the MNAR condition of their study. It is worth noting that in their MNAR condition, lower-proficiency examinees (i.e., those with θ < 0) were simulated to have virtually always (i.e., 95% of the time) chosen the harder of the choice items. While still an important and unique study, this particular condition may be unrealistic if proficiency is posited to give the examinee some greater insight into true item difficulty. Rather than being almost always attracted to the harder item, the lower proficiency examinees may more likely be unable to distinguish between the item difficulties. Therefore, these examinees may arbitrarily choose between items, or choose based on some item characteristic other than difficulty, such as order of presentation of the items.

In order to study examinees' behavior under examinee-selected MC item designs, Wang et al. (1995) administered 20 multiple choice Advanced Placement chemistry items with instructions for examinees to indicate their preferred item for three pairs. In particular, while Item 12 was substantially more difficult (with difficulty estimated from the operational administration under which choice was not given) relative to Item 11, those who indicated a preference for Item 12 in fact performed better on Item 11. While this provides some evidence that examinees may not make the best choice with respect to selecting the item that yields the highest score, it is unclear whether examinees may (a) fail to discern relative item difficulty; (b) lack insight into their own proficiency; or (c) choose items for reasons other than maximizing their expected score. What must be determined is whether the examinee item selection mechanism is affected by the item, rater, or examinee characteristics to be estimated.

Possible Theory Behind How Examinees Select Items. While proponents of examinee-selected items posit that examinees understand what they know best and therefore choose the item that will maximize their final score, the evidence is unclear. As Wainer and Thissen (1994) note, the choice is, at best, made based on examinees' subjective probabilities of earning each possible score on each item. Therefore, it comes with all of the known problems that human subjects have with estimating probability under uncertainty (Tversky & Kahneman, 1974). As Kruger and Dunning (1999) note, people tend to overestimate their own proficiency, which could lead them to make poor choices, despite perhaps having the rational goal of maximizing their expected test scores. If examinees are not, in fact, reliable judges of their own proficiency, other factors may influence the test items they choose.

Rather than subject area proficiency, it may be that a characteristic independent of proficiency affects examinees' subjective probability estimates and therefore the selection of items. One likely candidate for such a trait, that is *a priori* independent from the examinee's proficiency, is termed "test wisdom." Millman, Bishop, and Ebel (1965) define test wisdom as "a

subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score" (p. 707). If certain items lend themselves to being graded more leniently and if test-wise examinees can discern that item trait, then test wisdom may indeed be an important aspect of item selection.

Millman et al.'s (1965) outline of traits possessed by test-wise individuals focuses on multiple choice (i.e., objective or selected response) items, but the general principles of time management, error avoidance, and the consideration of the test publisher's intent apply more generally to other item formats as well (Sarnacki, 1979). Generalizing these principles to examinee-selected CR items, test-wise individuals may be able to determine which item is the least demanding, most time-efficient way to generate a comprehensive response that is expected to yield the maximum item score. Test-wise examinees may better avoid going off-topic through an understanding of what the test publisher aims to assess. Sarnacki (1979) argued that examinees who possess test wisdom "experience a general sense of security in taking tests" (p. 263), and are therefore in a better position that an examinee whose content area knowledge is similar, but who lacks test wisdom.

Test wisdom and proficiency may be conflated when examinee proficiency itself affects item selection. Chi (2006) outlines seven major aspects that enable expert (i.e., high proficiency) individuals to excel in their respective fields. Two of those aspects that are particularly relevant to examinee item selection are (a) detection and recognition, wherein experts are better able to detect difference in the items; and (b) opportunism, in that experts tend to make use of whatever resources are available. For example, expert physics students tended to rely more on the underlying physical principle and less on superficial non-physics problem characteristics in judging problem difficulty and were therefore better able to detect differences in problem difficulty (Chi, Glaser, & Rees, 1982, p. 65-68). If greater proficiency on the construct being measured better enables examinees to identify the item that is easiest, then

it is important to consider situations in which examinees' choices depend upon the latent proficiency trait (θ).

2.2. Relevant Models for Constructed Response Items

Single Continuous Latent Trait Models. Before outlining the methods to be used in this study, it is helpful to review some commonly used models for CR items. Following Thissen and Steinberg's (1986) terminology, the graded-response model (GRM; Samejima, 1969) is a "difference model" or what Embretson and Reise (2000) term an indirect model, appropriate for the sort of ordered response category data that the rating of CR items generate. Specifically, the model may be represented with a logistic link function as follows:

$$P(Y_i = k + 1 \mid \theta) - P(Y_i = k \mid \theta) = \frac{1}{1 + e^{(\lambda_i \theta - \zeta_{ik})}} \quad (1)$$

where Y_i designates the polytomous essay score for item i, k indexes score categories from 1 to K, θ is examinees' continuous latent proficiency trait, ζ_{ik} are the K - 1 item thresholds for item i, and λ_i is the discrimination parameter for item i. By definition, $P(Y_i = 0 \mid \theta) \equiv 0$.

An alternative specification of a polytomous IRT model is the generalized partial credit model (GPC; Muraki, 1992). It differs primarily in that it is a "divide-by-total" model—what Embretson and Reise (2000) would call a direct model—and is also appropriate for ordered polytomous responses of the sort that arise from the rating of constructed responses. Also using a logistic link function, this model may be represented as follows:

$$P(Y_i = k \mid \theta) = \frac{e^{\sum_{m=0}^{k} (\lambda_i \theta - \zeta_{im})}}{\sum_{v=0}^{K-1} e^{\sum_{g=0}^{v} (\lambda_i \theta - \zeta_{ig})}} \quad (2)$$

where Y_i designates the polytomous essay score for item i, k indexes score categories from 1 to K, θ is examinees' continuous latent proficiency trait, ζ_{ik} are the K - 1 item thresholds for item i, and λ_i is the discrimination parameter for item i. By definition, $\sum_{g=0}^{0} (\lambda_i \theta - \zeta_{ig}) \equiv 0$.

These two models characterize the items as direct indicators of examinee proficiency, without formally modeling any effects of the human raters who generate the item scores. Many studies of CR item choice acknowledge the additional complexity of analyzing scores on CR items that were scored by human raters (Allen, Holland, & Thayer, 2005). Understanding that raters may vary in their severity and variability of ratings, Patz, Junker, Johnson, and Mariano (2002) formalized the hierarchical rater model (HRM) introduced by Patz (1996):

$$P(Y_{ij} = k \mid \eta_i = \eta) \propto e^{\left\{-\frac{1}{2\psi_{ij}^2}\left[k - (\eta - \varphi_{ij})\right]^2\right\}} \quad (3)$$

$$P(\eta_i = \eta \mid \theta) = \frac{e^{\sum_{m=0}^{\eta}((\theta - \beta_i) - \gamma_{im})}}{\sum_{v=0}^{K-1} e^{\sum_{g=0}^{v}[(\theta - \beta_i) - \gamma_{ig}]}} \quad (4)$$

where Y_{ij} is rater *j*'s rating on a 1-to-*K* discrete scale of the examinee's response to item *i*; η_i is the examinee's latent class membership for item *i*; ψ_{ij}^2 and φ_{ij} are variance and rater severity parameters, respectively, associated with rater *j* and item *i*; θ is the examinee's latent continuous proficiency; β_i is the location parameter for item *i*; and γ_{ik} are the *K* - 1 step parameters for item *i*. Note that Expression (3) follows DeCarlo et al.'s (2011) representation of Patz et al.'s (2002) Expression (5), which specifies φ_{ij} as a severity parameter, rather than a leniency parameter.

Multiple ratings per constructed response allow for the estimation of individual rater characteristics (i.e., discrimination and location of latent thresholds) and therefore provide much richer information than simple inter-rater agreement or other aggregate statistics. It is critically important for the dependency among different raters' scoring of the same examinee's work to be properly modeled. In Patz et al.'s (2002) HRM, ratings of examinee work are direct indicators of latent essay quality—specified as categorical—and the essay qualities are direct indicators of examinee proficiency—specified as continuous. While the previous IRT-based approaches represent one means of modeling CR items, signal detection theory (SDT; Wickens, 2002) represents an alternative approach and the HRM (Patz et al., 2002) uses SDT-like parameters.

Some Background on Signal Detection Theory (SDT). In a very general sense, SDT models may profitably be applied in situations when human subjects are asked to judge whether a condition is present. It has been often applied in psychophysics, for example, to model subjects' perceptions of visual or auditory stimuli. Examples where SDT models have been successfully employed include experiments when subjects are asked to judge whether an auditory tone is present or to identify whether a given word was on a list of words to be memorized (Wickens, 2002). The basic model has two components: the discrimination with which subjects detect conditions and the level of perception required to conclude that a condition is present. Just as subjects are hypothesized to have an internal (i.e., latent) criterion for determining if a condition is present, so too may exam raters have latent criteria for whether a given item response demonstrates subject-area mastery.

Figure 1 visually depicts a hypothetical signal detection task. In this case, raters must make some latent classification (e.g., essay score category) on an ordinal, four-point scale (e.g., as in a holistic essay rubric). This figure shows that raters' perceptions—represented by the horizontal axis—vary continuously for the phenomena that they are asked to judge. Consider a situation in which a rater is presented with a stimulus whose true latent classification is 2 on a 1-to-4 scale. The conditional distribution of the rater's perception, given that true latent classification, is shown in gray (Figure 1). This figure demonstrates that the rater is likely to give a rating of 2—since most of the mass of the conditional distribution lies between his first and second thresholds (c_1 and c_2). However, note the substantial remaining mass for this conditional distribution lies either below c_1 or above c_2, which correspond to incorrect classifications: categories 1 and 3, respectively. Such misclassifications result from rater perceptions either being too high or too low. With greater values of d (i.e., rater discrimination), the conditional probabilities associated with the latent classes become increasingly separated and raters

classify more phenomena correctly. This framework lends itself naturally to the classification task of detecting the level of mastery as demonstrated by an item response.

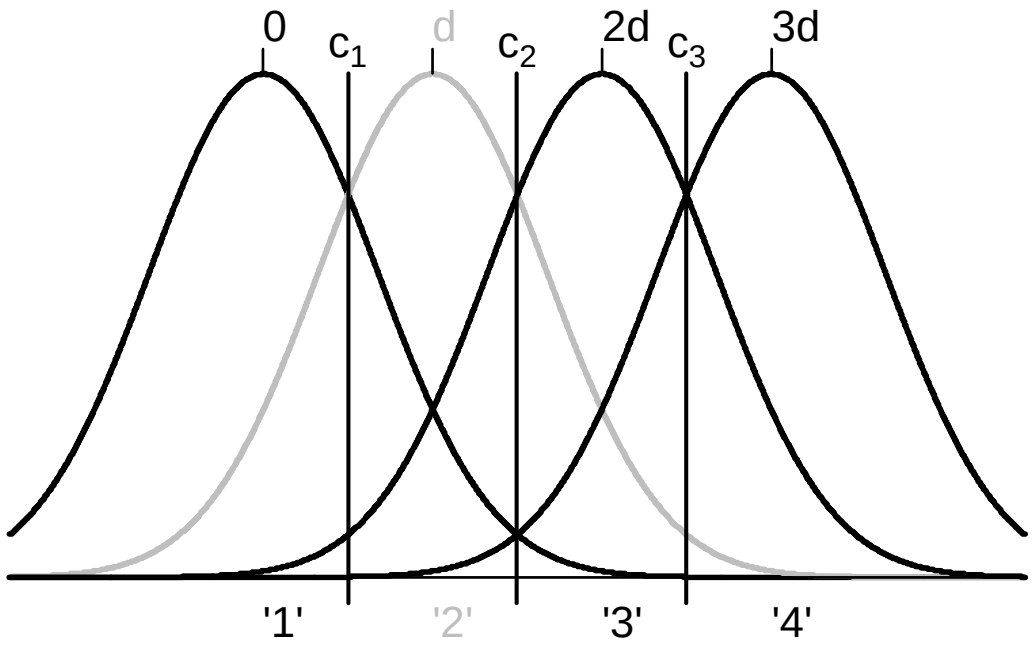

Figure 1. Graphical depiction of hypothetical signal detection task.

Signal Detection as a Rater Model. Rather than treating the constructed responses as direct indicators of a continuous latent trait (as the previously cited polytomous IRT models do), it may be more appropriate to treat them as categorical. DeCarlo (2002) demonstrated how a SDT framework may be applied, treating the rating task as one of latent classification. Such a model requires positing that raters, rather than directly judging the continuous latent proficiency of the examinee as does IRT, instead rate essays on an ordered, categorical scale.

$$P(Y_{ij} = k \mid \eta_i = \eta) = F(c_{ijk} - d_{ij}\eta_i) \quad (5)$$

where Y_{ij} is rater *j*'s rating of the examinee's response to item *i*; $F(\cdot)$ is an arbitrary, cumulative link function (e.g., the cumulative logistic function); and d_{ij} and c_{ijk} are rater *j*'s discrimination and K - 1 location thresholds, respectively; η_i is the examinee's latent class membership for item *i*.

This model allows for allows for flexibility, permitting raters' latent criteria to approach positive or negative infinity, accounting for the fact that raters may not use the full scale of rating categories. In particular, since the raters score items on a discrete, ordinal scale (e.g., from 1 to 4 as in Figure 1) and given that the true class of the essay is not known, the raters' task is to classify works onto the scale defined by the holistic rubric. Such classification can naturally be modeled using latent class analysis. This differs somewhat from traditional polytomous IRT approaches, which presuppose that examinee proficiency is judged directly by the rater and that the essay quality may vary continuously.

Indeed, there are a variety of rater effects that could not be adequately characterized under Patz et al.'s (2002) formulation of the HRM. Raters can vary on a wide variety of important characteristics, such as a tendency to only utilize some part of the rating scale or a tendency to give ratings near the center of the scale (Myford & Wolfe, 2003, 2004). DeCarlo (2002) provided a rater model that is flexible enough to detect these more nuanced rater effects and laid the groundwork for a re-conceptualization of the HRM.

Latent Class Rater Model with Single Continuous Latent Trait. In settings where multiple raters have rated multiple essays examining a common construct, DeCarlo's (2002) rater model may be generalized. DeCarlo, Kim, and Johnson (2011) re-conceptualized the HRM (Patz, 1996) with discrete, ordinal latent classification of essay quality as direct indicators of a continuous, latent trait (i.e., proficiency). Figure 2 shows a structural equation modeling representation of DeCarlo et al.'s (2011) model. The first level of the model is the SDT rater portion, where raters judge the quality of essays produced by examinees and map those perceptions onto the rating scale—see the mapping of the raters' perceptions (Ψ_{ij}) onto item ratings (Y_{ij}). The second level is an IRT-like model where latent rater classifications are related to examinee proficiency. This level is represented in Figure 2 as the mapping of the examinee proficiency (θ) onto the true latent class membership for each item (η_i). This structure explicitly separates the characterization of the essay-rating task from the estimation of examinee proficiency. In other words, the item and rater parameters are separate and estimable under the model. Such a model follows naturally from the latent classification task implicit in the use of scoring rubrics with fixed, discrete rating scales and appropriately treats the dependency among multiple ratings of the same essay. In using such a model, rater characteristics may be estimated, in addition to the item and examinee characteristics that may be estimated using other approaches.

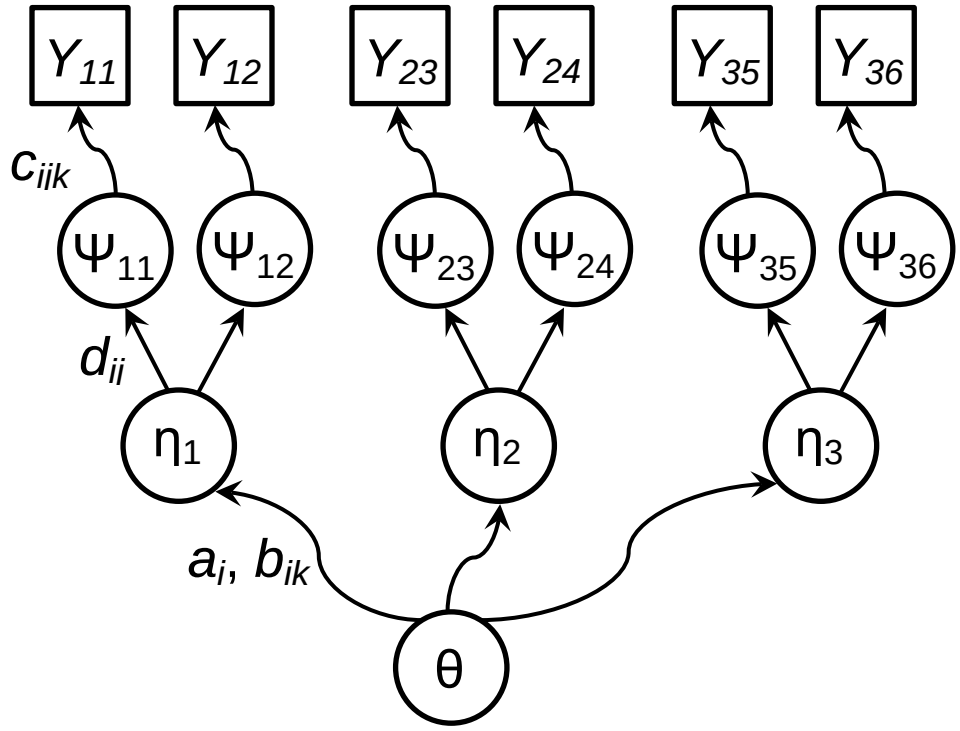

Figure 2. Structural equation model (SEM) representation of an HRM-SDT for two raters, each (six total) having rated constructed responses per examinee.

The HRM-SDT may be described by the Level 1 [i.e., rater level; see Expression (6)] and the Level 2 [i.e., item level; see Expression (7)] adapted from DeCarlo, et al. (2011):

$$P(Y_{ij} = k \mid \eta_i = \eta) = F(c_{ijk} - d_{ij}\eta_i) \tag{6}$$

$$\log\left[\frac{P(\eta_i = \eta + 1 \mid \theta)}{P(\eta_i = \eta \mid \theta)}\right] = a_i\theta - b_{ik} \tag{7}$$

where Y_{ij} is rater j's rating of the examinee's response to item i; $F(\cdot)$ is an arbitrary, cumulative link function (in this case the cumulative logistic function); d_{ij} and c_{ijk} are rater j's discrimination and K - 1 location thresholds, respectively; η_i is the examinee's latent class membership for item i; θ is examinee proficiency; and a_i and b_{ik} are item discrimination and K - 1 location thresholds, respectively. To simplify notation, the number of latent classes and the number of essay rating categories were both fixed to K, but this need not be the case in general. The rater level is a

latent class, signal detection theory model, with the item level being modeled as a generalized partial credit IRT model (GPC; Muraki, 1992). Rewriting the item level model, Expression (7), in probability terms, yields:

$$P(\eta_i = \eta \mid \theta) = \frac{e^{\sum_{m=0}^{\eta} (a_i\theta - b_{im})}}{\sum_{v=0}^{K-1} e^{\sum_{g=0}^{v} (a_i\theta - b_{ig})}} \tag{8}$$

With a few simplifying assumptions (DeCarlo et al., 2011), the unconditional probability for observing an arbitrary set of constructed response ratings for the HRM-SDT follows:

$$P(\boldsymbol{Y}) = \sum_{\eta} \left[\prod_{ij} P(Y_{ij}|\boldsymbol{\eta})\right] \int_{\theta} \left[\prod_{i} P(\eta_i|\theta)\right] P(\theta)\, d\theta \tag{9}$$

where ***Y*** is the pattern of constructed response ratings for an examinee across all *J* raters and all *I* items and **η** is the pattern of true latent classifications of constructed responses across all *I* items.

2.3. Some Missing Data Terminology

When test publishers implement examinee-selected item designs, they introduce missing data by design into the assessment. In other words, when an examinee chooses to answer Item 1 rather than Item 2, their response to Item 2 is not observed: that datum is, in effect, missing. So the missing data mechanism in this context is the examinee's choice of which item to answer. Consequently, it will be helpful to give an overview of some missing data terminology. This study will characterize the problem of incomplete data that arises from assessments allowing examinees some measure of choice using the widely utilized missing data nomenclature given by Little and Rubin (2002).

Missing data arises from what is called a mechanism for missing data, or the theoretical process by which observed data arise with some level of missingness. That mechanism may be characterized in one of three general classes (a) missing completely at random (MCAR); (b) missing at random (MAR); and (c) missing not at random (MNAR, also not missing at random or

NMAR). When data are MCAR, the missing data mechanism depends neither on the observed data, nor on the unobserved parameters that one wishes to estimate. Consider a full dataset where 10% of observed values on a particular are arbitrarily deleted (i.e., set to missing); that is an example of data being missing completely at random. Data that arise from a missing data mechanism that depends solely on the observed data are called MAR and missing data mechanisms that depend on the unobserved values or parameters are MNAR. An example of MAR data in this context would be if there were an observed covariate, such as gender, that explained the missing data. In such a situation, an examinee's gender—a known feature of examinees—would determine whether the data were missing or present. Missing data mechanisms that depend on the unobserved values or parameters are MNAR. The data would be MNAR if examinees' item selection were related to some unobserved characteristic that appears in the model to be estimated. For example, if latent proficiency (θ) affected item selection, then the data would be MNAR, since θ is an unobserved, examinee-level trait to be estimated by the model.

Another important assumption underlying most psychometric models and the statistical packages used to estimate them is the notion of ignorability. The likelihood function may be written in two parts: one, which characterizes the missing data mechanism, and another which characterizes the model of primary interest. When making likelihood-based inferences in either the maximum likelihood (ML) or Bayesian modeling frameworks, if data are MCAR then the likelihood function ignoring (i.e., excluding from estimation) the missing data mechanism may be used to estimate the desired parameters, rather than the more complicated full likelihood function that incorporates the missing data mechanism. If the data are MAR and the parameters that give rise to the missing data are distinct from those that are being estimated, then the likelihood function that ignores the missing data mechanism may also be used to estimate the parameters of interest without introducing bias (Rubin, 1976). In this context, if the provision of

examinee choice in which items to respond to (i.e., the missing data mechanism) may be ignored when estimating the parameters of interest (i.e., examinee, item, and rater characteristics), then that estimation is a great deal simpler, so ignorability is desirable.

Some studies have tested the assumption that examinee item selection is ignorable for the estimation of examinee proficiency. Using data collected by Bridgeman, et al. (1997), Allen, Holland, and Thayer (2005) examined whether estimating the standard (i.e., simpler) IRT model that does not explicitly model the missing data mechanism (i.e., the examinee item selection process) is appropriate. Such investigation aimed to shed light on whether assuming ignorability was reasonable. The authors found that there were significant ($p < .05$), positive relationships between choosing a given essay topic and the score earned on that topic in five of eight samples, with no significant relationship in the remaining three samples. So in more than half of the samples, the better examinees performed on Topic A, the more likely they were to have identified it as a preferred item. Inversely, there was a significant ($p < .05$), negative relationship between preference and the score earned on the non-preferred topic in four of eight samples, with no significant relationship in the remaining four samples. In other words, in half of the samples when examinees preferred Topic A, they tended to perform less well on Topic B than on Topic A. While not necessarily evidence for or against the appropriateness of assuming ignorability, these results should lead researchers to question the common assumption of ignorability. Wainer and Thissen (1994) noted an important question that remained at the time of their writing and that still appears to remain is, “How far from ignorable can nonresponse be and still be acceptably adjusted for statistically?” (p. 190). This research aims to shed light on that question in the context of the HRM-SDT.

2.4. Models Incorporating Examinee-Item Selection

Applying any of the previously discussed IRT models to data generated under examinee-selected item conditions has proved problematic (e.g., Bridgeman et al., 1997). Existing models for polytomous items with missing data as in Holman and Glas (2005) have been proposed and a few have been presented in the context of examinee-selected items. Two fundamentally different approaches to this problem exist: the first models the examinee-item selection as directly related to examinee proficiency (i.e., θ, or the underlying trait to be measured by the test), while the second incorporates a separate latent variable, representing examinee-item selection as different underlying trait that may be independent from or correlated with proficiency.

Lukhele et al. (1994) took the first approach and modeled the examinee-item selection process as directly related to proficiency. Analyzing the 1989 version of the Advanced Placement Exam in chemistry exam, they used a two-parameter logistic IRT model to model examinees' selection of one item from a pair of possible items and a two-parameter version of Bock's (1972) nominal response model for the selection of the set of three items chosen from a possible five prompts, with each of the ten unique combinations representing nominal responses. The multiple choice items were modeled under the three-parameter logistic IRT model (Birnbaum, 1968) and the constructed response items were modeled using Samejima's (1969) graded response model. Thus, Lukhele et al. (1994) presumed that the more proficient examinees would choose the "best" items—those they could answer more completely and efficiently—because of their superior knowledge of chemistry. Examinees with weaker chemistry skills are conversely assumed not only to perform more poorly on items, but to be less strategic in choosing which items to answer.

If one posits—as Lord (1983) did—that choosing items to answer is related not only to proficiency but a trait he called "temperament," (e.g., risk tolerance) then an alternate approach

is warranted. Modeling examinee-item selection as related to a latent variable that is unique to proficiency, Wang, Jin, Qiu, and Wang (2012) proposed a family of new models that are a type of multidimensional IRT model (Reckase, 2009). Their continuous, student test-wisdom parameter (γ) is added to proficiency (θ) when modeling polytomous item responses in order to free θ from the effect of the examinee selection of items and therefore ensure comparability of θ across examinees, regardless of their choice of items. This clearly represents a more flexible approach to the modeling of examinee-item selection than was used by Lukhele et al. (1994)—whose model is closely related to a special case of Wang et al.'s (2012) model in which γ and θ are restricted to being equal. The incorporation of examinee item selection into existing models, while relevant to contextualize this work, is beyond the scope of this study.

2.5. Current Study

The current study differs from existing studies of CR item models under an examinee-selected item design in a few important ways. Where much of the existing research on examinee-selected items utilizes IRT models for CR item ratings, this study will focus on the HRM-SDT (DeCarlo et al., 2011) for modeling CR item ratings. Since the HRM-SDT separates item and rater parameters, it will be possible to understand more clearly the recovery (i.e., good estimation) of item and rater parameters. For that reason, more comprehensive findings will be possible with respect to the effects of using examinee-selected items on examinee, item, and rater characteristics.

This study will simulate a number of possible manners in which examinees select an item to which to respond. This will enable a comparison of the performance of the HRM-SDT across those varied situations. While relatively few existing studies—Bradlow and Thomas (1998), being one—have specified possible item selection mechanisms, the simulations described herein will enable conclusions to be drawn with respect to the examinee item

selection processes that may impact the recovery of HRM-SDT item, rater, and examinee parameters. If the HRM-SDT exhibits good recovery of examinee, rater, and item parameters (i.e., with little bias and small RMSE) across the three examinee-item selection conditions, then it is appropriate for operational scoring of examinees' responses and monitoring of raters' performance in the context of tests with examinee-selected constructed response items.

Chapter III

METHODS

When examinees are instructed to choose to which of a subset of possible constructed response (CR) items to address, there may be impact—adverse or otherwise—on, for example, the prediction of their proficiency. In the context of the hierarchical rater model (HRM; Patz, Junker, Johnson, & Mariano, 2002) with signal detection theory (SDT) rater components (HRM-SDT; DeCarlo, Kim, & Johnson, 2011) the impact of using examinee-selected CR items should be considered for:

a) the bias and variance of estimates of the continuous latent trait or "proficiency" (θ);

b) classification in terms of the ordered latent classes (η_i); and

c) the bias and variance of estimates of item (a_i, b_{ik}) and rater (c_{ijk}, d_{ij}) parameters.

These issues will be investigated by simulating a number of replicates (i.e., independent datasets) in a variety of ways and attempting to recover examinee, item, and rater parameters.

3.1. Complete Data Generation

Assessment Design and Item Characteristics. These simulations were developed with the design of existing, high-profile assessments that employ examinee-selected CR items in mind. In particular, the AP Exam in United States (U.S.) history consists of a multiple choice (MC) section containing 80 items and a CR section which requires 3 essay responses. They have roughly one hour to complete the MC portion and approximately 2 hours to respond to the three CR items. All examinees must answer the first CR item. They must select one of either item 2 or 3 to respond to and must select one of either item 4 or 5 to complete the CR section (College Board, 2010a). The AP Exam in European history follows a similar structure, but rather than selecting two items from two sets of two possible essays, examinees must choose two items from two sets of three possible essays, for the same total of three essays—including the common first item (College Board, 2010b).

The hypothetical assessment from which data will be simulated for this study is a great deal simpler. No multiple choice items will be considered, though they may profitably be added to an assessment in the HRM-SDT context (Kim, 2009). Additionally, rather than selecting two items, from either two separate pairs (as in AP U.S. history) or two separate sets of three (as in AP European history), this hypothetical assessment will require that examinees choose from a pair of possible items and all will answer a common third item. In other words, examinees will choose to respond to either Item 1 or Item 2 and all will be required to respond to Item 3. Finally, rather than scoring the items on a nine-point scale as is done for the AP exams in question, the simulated data will be scored on a six-point scale. The data will be simulated with complete data following an underlying HRM-SDT (DeCarlo et al., 2011)

The population item parameters will also be fixed across conditions and replicates (see Table 2). In particular, that table shows that the two items under examinee item selection (i.e., Items 1 and 2) differ greatly in terms of their location (i.e., threshold) parameters (b_{ik}). This large

discrepancy on item location is included to represent two items that differ greatly in terms of difficulty, which—when administered as a set for examinee item selection—may hamper the precise and accurate recovery of item, examinee, and rater characteristics. In other words, this is a worst-case scenario with respect to the relative difficulty of examinee-selected items.

Table 2.

Population Item and Rater Parameters

	Item (*i*)		
Parameter	**1**	**2**	**3**
a_i	1.0	1.0	1.5
b_{i1}	-1.7	-0.7	-1.6
b_{i2}	-1.1	-0.1	-0.8
b_{i3}	-0.5	0.5	0.0
b_{i4}	0.1	1.1	0.8
b_{i5}	0.7	1.7	1.6

	Item (*i*)					
	1		2		3	
	Rater (*j*)					
Parameter	**1**	**2**	**3**	**4**	**5**	**6**
d_{ij}	2.2	3.8	1.8	4.2	2.0	4.0
c_{ij1}	1.1	1.9	0.9	2.1	1.0	2.0
c_{ij2}	3.3	5.7	2.7	6.3	3.0	6.0
c_{ij3}	5.5	9.5	4.5	10.5	5.0	10.0
c_{ij4}	7.7	13.3	6.3	14.7	7.0	14.0
c_{ij5}	9.9	17.1	8.1	18.9	9.0	18.0

Rater and Examinee Characteristics. In order to reflect the nature of how CR item ratings arise six simulated raters of varying discriminatory skill will be considered. The same two of six raters, will rate only one of the possible three CR items under consideration. Since one of the test items is required, only four total ratings will be observed for each examinee—two for the required item and two for the examinee-selected item. A variety of d_{ij} values were selected from the plausible range of 1 to 4 and c_{ijk}[1] were equally spaced and placed at the intersection of the adjacent conditional distributions of rater perception (DeCarlo, 2008). The population rater parameters will also be fixed across examinee item selection conditions and replicates and are given in Table 2.

For all 30 replicates, the responses of 5,000 examinees each will be simulated. The continuous, latent examinee trait representing innate content area knowledge (i.e., proficiency) will be simulated under the following distribution:

$$\theta \sim \text{Normal}(0, 1) \tag{10}$$

Based on θ and the population item parameters, examinees will also have a true latent classification on each of the three items (η_i) that follow from the HRM-SDT.

3.2. Examinee Item Selection

Inherent to the design of this hypothetical assessment is the use of examinee-selected items. By choosing Item 2, for example, the examinee implicitly introduces missingness into his or her test data for Item 1 and it is that missingness that is at the heart of this study. Following Little and Rubin's (2002) notation, M_i is used to denote whether item i is missing as a result of an examinee not selecting item i. In the example of an examinee selecting Item 2—and therefore not choosing Item 1—the missing data indicators would take on values of 1 for M_1 and 0 for M_2. There are a number of possible characteristics that could contribute to an examinee

[1] Rater parameters c_{ijk} and d_{ij} are fixed at the rater level and could therefore equally accurately have been denoted c_{jk} and d_j, but the i was retained for clarity.

selecting one item over another, but the focus of this study will be to consider three conditions that vary on how related proficiency (θ) and an independent latent characteristic are to examinees' item selection.

Condition 0: Full Data. Using the same 30 replicates for which varying item selection mechanisms were simulated, Condition 0 refers to the full data condition. In other words, under Condition 0, all examinees responded to all three items. This condition will provide additional context for the interpretation of the remaining conditions, in which examinees select from among Items 1 and 2 and therefore have incomplete data.

Condition 1: Random Item Selection. The best case with respect to missingness would be if the examinee-selected item responses were missing completely at random (MCAR). Examinee item selection Condition 1 (henceforth, simply Condition 1) specified that examinee item selection—and therefore missingness of item i, M_i—is completely random. More formally:

$$M_1 \sim \text{Bernoulli}(p) \tag{11}$$

$$M_2 = 1 - M_1 \tag{12}$$

where M_i is a dichotomous indicator of whether item i is missing for the examinee in question ($M_i = 1$), or whether it is observed ($M_i = 0$) and p is the population proportion of examinees for whom Item 1 is observed (i.e., who have selected Item 1 over Item 2). In other words, in Condition 1, the expected proportion choosing item 1 will be p and the expected proportion of examinees selecting Item 2 will be $(1 - p)$. For this study, p is set to ½, leading to equal expected response rates for Items 1 and 2.

Condition 2: Item Selection due to Test Wisdom. It has been suggested (e.g., by Millman et al., 1965) that some latent construct other than proficiency may grant so-called "test-wise" individuals greater insight into the difficulty of items. Considering the real possibility of this phenomenon affecting examinees' choices in the assessment framework, Condition 2 will specify that examinee item selection is determined by the presence of a latent dichotomous test-wisdom indicator, δ. More formally:

$$\delta \sim \text{Bernoulli}(\omega) \tag{13}$$

$$P\left(M_1 \mid \delta, \theta\right) = P\left(M_1 \mid \delta\right) \tag{14}$$

$$M_1 = \begin{cases} 0 & \mid \delta = 1 \\ \text{Bernoulli}(.5) & \mid \delta = 0 \end{cases} \tag{15}$$

where δ is the dichotomous indicator of test wisdom; ω is the population proportion of examinees who possess test wisdom (i.e., for whom δ = 1). In other words, in Condition 2, the expected proportion of examinees selecting Item 1 will be (½·ω + ½) and the expected proportion choosing Item 2 will be (½ - ½·ω). In this study, ω is fixed at ½, leading to expected response rates for Items 1 and 2 of ¾ and ¼, respectively. Since in Condition 2, examinee item selection is due to test wisdom, which is defined as independent of proficiency (θ), Expression (14) shows that the missing data mechanism is independent of examinee proficiency (θ).

Condition 3: Item Selection due to Proficiency. On the other hand, rather than test wisdom guiding examinees' item selection as is proposed in Condition 2, examinee proficiency may be related to item selection. There is evidence to suggest that higher proficiency examinees may be better capable of detecting the difficulty of items (Chi et al., 1982) and therefore that perhaps Expression (14) does not hold in all settings or for all examinee populations. If higher-proficiency examinees can compare the relative difficulty of the items subject to examinee selection, they may be more likely to select the easier item. Condition 3 simulates the missing data mechanism as being related to a critical threshold of proficiency: $\theta = 0$. In particular, examinees above that threshold always select the easier of the two items, and examinees below that threshold—perhaps unable to detect differences in item difficulty—choose randomly between the two items. As is shown in Table 2, the location parameters for Item 1 are less than those for Item 2, so those above the threshold for θ have $M_1 = 0$, while those below the threshold either choose the harder Item 2 and thus $M_1 = 1$ and $M_2 = 0$ or they choose the easier Item 1 and $M_1 = 0$ and $M_2 = 1$. Formally:

$$M_1 = \begin{cases} 0 & | \ \theta \geq 0 \\ \text{Bernoulli}(.5) & | \ \theta < 0 \end{cases} \tag{16}$$

In this study, the critical threshold for savvy item selection was set at $\theta = 0$, thus leading to expected response rates for Items 1 and 2 of ¾ and ¼, respectively.

3.3. Model Estimation

The missing data mechanism (via examinee item selection) in Condition 1 is what Little and Rubin (2002) call missing completely at random (MCAR) for the estimation of the HRM-SDT. Because the missingness is unrelated to either the observed or missing data or to the parameters to be estimated, the data are MCAR. Mathematically, the general missing data mechanism—$P(M_i \mid Y_{Obs}, Y_{Mis}, \theta, \delta, \eta_i, a_i, b_{ik})$—collapses to this simpler form: $P(M_i \mid p)$. Condition 1 is expected to represent a best-case scenario in terms of accurate and precise

parameter recovery, since there was no dependence between item selection and the underlying examinee characteristics. In particular, Wainer and Thissen (1994) showed in the IRT context that unbiased item threshold parameter estimates may only be obtained from a random sample of the examinee population. This condition, which could alternatively be thought of as random assignment of examinees to items, should therefore yield estimates with little bias.

In Conditions 2 and 3, however, the missing data mechanism depends to varying degrees upon unobserved parameters. Specifically, examinee item selection is related to either examinees' test wisdom (δ) in Condition 2 or their own proficiency (θ) in Condition 3, so the data are missing-not-at-random (MNAR). In Condition 2, the general missing data mechanism collapses to: $P(M_i \mid \delta, b_{ik})$; while in Condition 3, it has the form: $P(M_i \mid \theta, b_{ik})$. As such, the probability of observing an arbitrary vector of CR ratings may not be simplified to the extent possible for Condition 1. Therefore the likelihood function for such an arbitrary vector of CR ratings is non-ignorable for the estimation of examinee, item, and rater characteristics under Conditions 2 and 3.

In estimating the HRM-SDT, Latent GOLD presumes that any data are at least MAR, with distinctness among parameters to be estimated and those that cause missingness (Vermunt & Magidson, 2005). This two-part requirement—the "ignorability assumption" mentioned earlier—is fundamental to many estimation techniques. In other words, the algorithms in Latent GOLD implicitly assume that the models may be estimated based solely on the observed data, regardless of whether the user has investigated the appropriateness of such treatment.

Other features of Latent GOLD's estimation techniques are pertinent to the estimation of the HRM-SDT. The models will be estimated with 21 quadrature points for the continuous latent trait (i.e., proficiency or θ) and Bayes constants of 2 will be added for the latent classes and the continuous latent trait in order to prevent boundary solutions (Vermunt & Magidson, 2008). The

use of Bayes constants does introduce bias, but it has been shown to more effectively avoid boundary estimation problems for the latent class component of the HRM-SDT than does maximum likelihood (Galindo Garre & Vermunt, 2006). The increase in bias is expected to be relatively small given the large sample (N = 5,000), which will contribute substantially more to the updating of the posterior than the prior. Despite possible boundary solution issues that may arise, maximum likelihood estimation will be used as a basis for comparison against posterior mode estimation.

3.4. Model Comparison

The results of the simulations must be compared in order to determine what, if any, impact the use of examinee-selected items has on the estimation of the HRM-SDT. The key statistics that will be estimated for comparison across conditions are:

1. bias and root mean squared-error (RMSE) of estimates of θ;
2. bias and RMSE of estimates of θ by known θ;
3. latent class recovery;
4. bias and standard errors of estimates of item parameters (a_i, b_{ik}); and
5. bias and standard errors of estimates of rater parameters (c_{ijk}, d_{ij}).

Comparison across conditions of each of these statistics will give insight into the extent to which the varying degrees of violation of ignorability affect model performance and will be discussed in turn. The performance of the latent class components of the model will also be evaluated using weighted kappa (Cicchetti & Allison, 1971) and percent exact agreement of the true and estimated latent classes.

Estimated bias and RMSE of the recovered θ will be examined in a variety of ways. First, the bias and RMSE will be plotted against the true θ for each condition in a single figure. Next, considering that bias and RMSE may vary in certain conditions depending upon the item

that examinees choose to respond to (i.e., M_1), additional plots will be produced by condition, separating out examinees who answered the easier item ($M_1 = 0$) from those answering the harder item ($M_1 = 1$). Similarly, bias and RMSE—conditional on true proficiency—were investigated by the components of the item selection mechanism. That is, conditional bias and RMSE were analyzed by δ for Condition 2 and on either side of the threshold for θ (i.e., 0) that governed selected item.

To get a sense of the impact on the recovery of item and rater characteristics, results must be considered across examinee item selection conditions. It is expected that the use of examinee-selected items will not unduly affect the recovery of item or rater characteristics in Condition 1, since examinee item selection in that condition is independent of examinee, item, and rater characteristics. And since the third item is required by all examinees, it is hypothesized that any effects of examinee item selection will be less severe for Item 3, relative to either Items 1 or 2, or Raters 5 and 6, relative to Raters 1 through 4. In particular, the bias and RMSE of the item discrimination parameters (a_i), item thresholds (b_{ik}), rater discrimination parameters (d_{ij}), and rater thresholds (c_{ijk}) will be examined for each condition, across replicates.

Chapter IV

RESULTS

In order to evaluate how well or poorly the population model was recovered under the examinee-item selection conditions, a few standard statistics were used. The bias of an estimate was calculated as the difference between the estimated parameter and its population value and that difference was averaged across replicates. For examinee-level parameters, the bias and squared deviation were calculated across all examinees and replicates. This gives a sense of both the direction and magnitude of any possible bias. However, it is also helpful to have an indication of bias that is consistent across parameter scales and that ignores the direction of the bias. Consequently, the absolute value of the bias, divided by the relevant population value, multiplied by 100 was computed to yield the absolute percent bias. Additionally, to understand the variability of the model estimates the root-mean squared-error (RMSE) was estimated as the square root of the mean squared deviation across replicates of the parameter estimate from its population value.

To better contextualize these statistics, this section will rely on the rules of thumb for classifying the absolute percent bias used by Flora and Curran (2004). In particular, absolute bias values less than 5% were considered negligible, values between 5% and 10% were moderate, and any value greater than 10% absolute bias was considered to indicate large bias. Section 4.1 discusses the recovery of examinee level categorical and continuous trait values, while Sections 4.2 and 4.3 focus on item and rater parameter recovery, respectively.

The use of posterior mode estimation with Bayes constants of 2 for the latent categorical and continuous examinee characteristics led to solutions that converged, without any boundary parameter estimates. For comparison's sake, the same analyses were performed under maximum likelihood; in the three conditions in which examinees selected from among two items (i.e., Conditions 1 through 3), a number of problems occurred. Three of thirty replicates under

Condition 1 exhibited boundary solutions. Eight such problems surfaced under Condition 2, along with four cases of failed convergence. And a dismal 21 of 30 boundary solutions were encountered for Condition 3 under maximum likelihood estimation. These findings reaffirm Galindo Garre and Vermunt's (2006) findings that posterior mode estimation leads to better estimation than maximum likelihood for models of this variety. An alternative approach to simply taking the mode of the posterior parameter density, would be to sample a number of draws from the posterior for each replicate. The computational intensity of such an approach for this complex model prevented its implementation, but could be investigated as a possible means for improving parameter recovery.

4.1. Examinee Parameter Recovery

Recovery of Proficiency. As the primary focus of this study was to investigate the effects of allowing examinees their choice of essay topics, it is natural that the recovery of examinee traits is of primary interest. The figures and tables that follow summarize the recovery of proficiency estimates in terms of bias and RMSE. The lower the magnitude of bias of the proficiency estimates, the closer the estimates are to the true values. The lower the RMSE, the less the model estimates vary around the true value.

Since there are some competing models that may be implemented in the case of constructed response scoring, a naïve polytomous IRT model in the form of the graded-response model (GRM; Samejima, 1969) was estimated. It is “naïve” in the sense that it ignores the natural dependence structure between the pair of ratings that two raters gave for the same item response. The deviation of proficiency estimates from this alternate model may not truly be called “bias”, as the IRT model was not the underlying population model that was used in simulating these data. Nevertheless, for the sake of parsimony, the term bias will be used for deviations of proficiency estimates from both the HRM-SDT and naïve IRT models.

A comparison of estimated and true proficiency for both models is of primary interest. When analyzing the recovery of examinee proficiency, the bias (i.e., deviation of the posterior mode estimate from the true value) and that deviation squared were calculated at the examinee level. Because the pattern of results varied by true proficiency and in order to make clear inferences about these statistics, the conditional bias and RMSE were estimated within 0.5 unit intervals (i.e., “bins”). Thus the bias and RMSE of proficiency estimates are conditional upon the true proficiency values. Note that the following review of a single replicate is meant as a preliminary review of possible patterns that is not practical for all 30 replicates. While more conclusive inferences may be drawn from the results across all replicates, this deeper review of a single replicate revealed patterns that would be hard to detect across all replicates.

Figure 3 shows the estimated proficiency from the HRM-SDT and the naïve IRT model for the first replicate are plotted against examinees’ true proficiency under Condition 0, in which examinees responded to all items. Figure 4, Figure 5, and Figure 6 show that same information for Condition 1 (random item selection), Condition 2 (test-wise item selection), and Condition 3 (item selection due to proficiency). One feature that distinguishes Condition 0 (i.e., Figure 3) from similar plots for the remaining three conditions is that the estimates vary less around the true proficiency values. In other words, they are more tightly clustered around the solid diagonal reference line that corresponds to perfect recovery. Another differentiating feature is that for both the HRM-SDT and the naïve IRT model, the range of recovered proficiency estimates is slightly larger when examinees responded to all items (i.e., Condition 0) than when they selected either Item 1 or Item 2 (i.e., Conditions 1 through 3). The one commonality across the four plots is that the estimates tend to be slightly more compressed for the HRM-SDT than for the naïve IRT model.

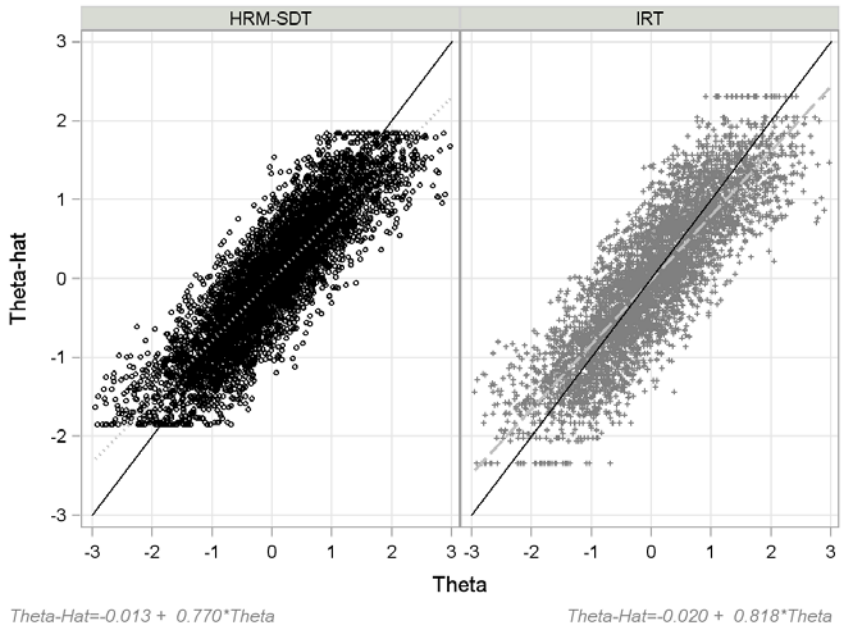

Figure 3. Proficiency estimates ($\hat{\theta}$) from the HRM-SDT and an IRT model by true proficiency (θ) for Condition 0 (full data, i.e., no item selection), under posterior mode estimation for replicate 1.

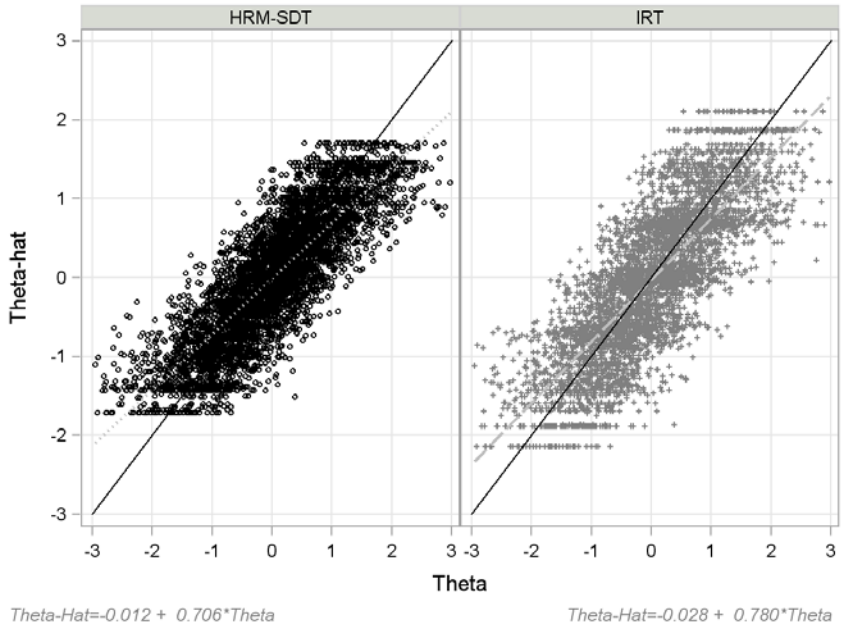

Figure 4. Proficiency estimates ($\hat{\theta}$) from the HRM-SDT and an IRT model by true proficiency (θ) for Condition 1 (random item selection), under posterior mode estimation for replicate 1.

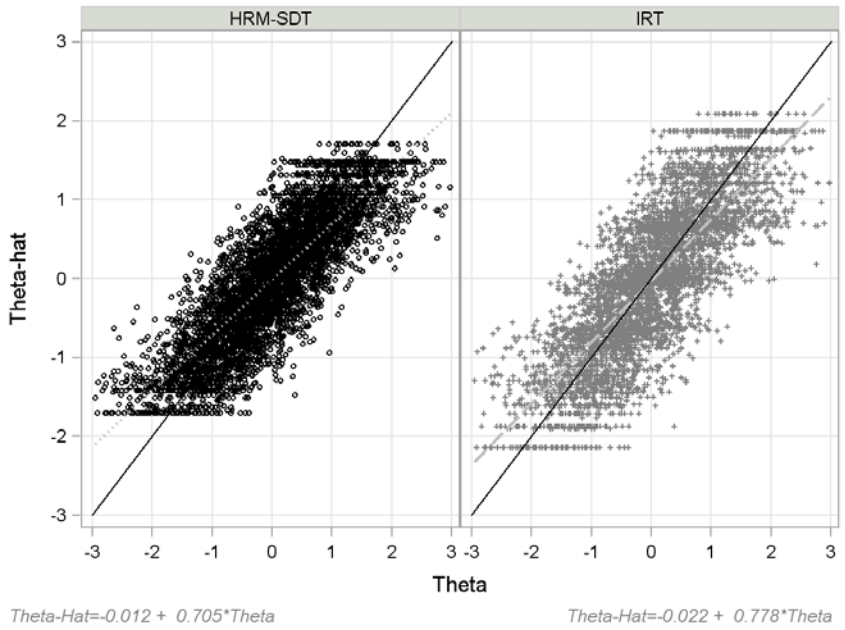

Figure 5. Proficiency estimates ($\hat{\theta}$) from the HRM-SDT and an IRT model by true proficiency (θ) for Condition 2 (test-wise item selection), under posterior mode estimation for replicate 1.

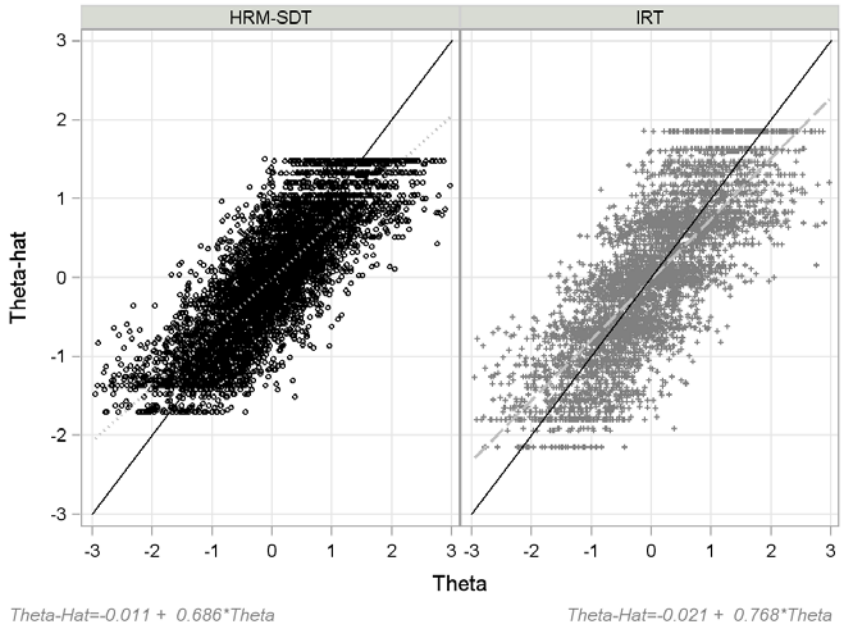

Figure 6. Proficiency estimates ($\hat{\theta}$) from the HRM-SDT and an IRT model by true proficiency (θ) for Condition 3 (θ threshold item selection), under posterior mode estimation for replicate 1.

The impact of the three possible examinee-item selection mechanisms—Conditions 1, 2, and 3—on the recovery of examinee proficiency is of paramount importance to judge the appropriateness of using these models under examinee-selected item designs. The similarity between Conditions 1 and 2—where examinee item selection is unrelated to proficiency—is apparent in the similarity of the plots of estimated by true proficiency in Figure 4 and Figure 5. Figure 6, on the other hand, shows greater deviation of the estimated from the true proficiency values; most notably, with relatively more estimates than were simulated at the higher end of the proficiency scale.

After averaging across all 30 replicates, the conditional bias of proficiency estimates from the HRM-SDT and naïve IRT model for the four conditions each are presented graphically in Figure 7 and also numerically in Table 3. The pattern of bias across the range of true proficiency is similar for both models with estimates of proficiency being shrunk toward zero—e.g., larger positive bias for estimates whose true values are larger and negative. The bias in estimated proficiency tended to be slightly larger in absolute magnitude (i.e., further from zero) for the HRM-SDT than the naïve IRT model across the range of true proficiency. The patterns of bias were fairly consistent comparing the same model across the four conditions, with the exception that the full data condition (i.e., Condition 0) tended to exhibit the smallest magnitude of bias across the proficiency range. The correlation of true proficiency with the estimates from the HRM-SDT was higher for all conditions than that with the estimates from the naïve IRT model, as the HRM-SDT was the population model underlying the simulated data. In practical terms, the majority of examinees are expected to be located between -2 and 2, the bias in the estimated proficiency for the most extreme condition (i.e., 3) ranged from -0.635 to 0.581 for the naïve IRT model and from -0.799 to 0.738 for the HRM-SDT.

Table 3.

Recovery of Examinee Proficiency by True Proficiency, Model and Condition

		Condition 0				Condition 1			
		Corr(θ, θ_{HRM})		Corr(θ, θ_{IRT})		Corr(θ, θ_{HRM})		Corr(θ, θ_{IRT})	
		.874		.857		.839		.809	
		Bias		RMSE		Bias		RMSE	
θ	*M*(*n*)	HRM	IRT	HRM	IRT	HRM	IRT	HRM	IRT
-3.0	12.700	1.346	1.202	1.371	1.269	1.553	1.355	1.576	1.415
-2.5	47.700	0.929	0.820	0.975	0.921	1.110	0.941	1.149	1.039
-2.0	138.300	0.580	0.494	0.670	0.657	0.717	0.576	0.790	0.740
-1.5	326.467	0.300	0.229	0.488	0.511	0.398	0.278	0.552	0.576
-1.0	598.233	0.131	0.078	0.441	0.484	0.180	0.100	0.474	0.560
-0.5	872.567	0.047	0.015	0.430	0.486	0.064	0.022	0.476	0.575
0.0	994.733	-0.003	-0.003	0.428	0.492	-0.002	-0.005	0.479	0.582
0.5	875.367	-0.050	-0.019	0.431	0.488	-0.067	-0.028	0.477	0.576
1.0	604.767	-0.132	-0.077	0.443	0.486	-0.181	-0.099	0.475	0.561
1.5	326.633	-0.303	-0.239	0.491	0.517	-0.404	-0.291	0.561	0.588
2.0	139.400	-0.581	-0.503	0.674	0.669	-0.723	-0.588	0.798	0.753
2.5	45.700	-0.923	-0.821	0.969	0.923	-1.100	-0.943	1.138	1.037
3.0	12.200	-1.359	-1.272	1.385	1.337	-1.548	-1.406	1.574	1.471
		Condition 2				Condition 3			
		Corr(θ, θ_{HRM})		Corr(θ, θ_{IRT})		Corr(θ, θ_{HRM})		Corr(θ, θ_{IRT})	
		.839		.810		.824		.799	
		Bias		RMSE		Bias		RMSE	
θ	*M*(*n*)	HRM	IRT	HRM	IRT	HRM	IRT	HRM	IRT
-3.0	12.700	1.490	1.298	1.514	1.366	1.564	1.341	1.591	1.406
-2.5	47.700	1.068	0.920	1.110	1.022	1.131	0.950	1.173	1.050
-2.0	138.300	0.684	0.557	0.762	0.731	0.738	0.581	0.815	0.754
-1.5	326.467	0.375	0.268	0.541	0.577	0.411	0.281	0.571	0.596
-1.0	598.233	0.180	0.103	0.479	0.561	0.196	0.108	0.495	0.586
-0.5	872.567	0.072	0.028	0.476	0.573	0.083	0.034	0.502	0.602
0.0	994.733	0.007	0.002	0.482	0.582	0.011	0.008	0.500	0.597
0.5	875.367	-0.057	-0.023	0.475	0.575	-0.071	-0.025	0.483	0.578
1.0	604.767	-0.181	-0.100	0.473	0.559	-0.198	-0.106	0.482	0.560
1.5	326.633	-0.413	-0.296	0.562	0.587	-0.450	-0.316	0.586	0.591
2.0	139.400	-0.748	-0.604	0.816	0.761	-0.799	-0.635	0.859	0.778
2.5	45.700	-1.131	-0.962	1.165	1.051	-1.192	-1.001	1.222	1.081
3.0	12.200	-1.603	-1.445	1.625	1.503	-1.668	-1.490	1.686	1.541

Note. RMSE = root-mean squared error. Posterior mode estimation results for 30 replicates. Corr(θ, θ_{HRM}) and Corr(θ, θ_{IRT}) are correlations of true proficiency with HRM-SDT and the IRT model estimates, respectively.

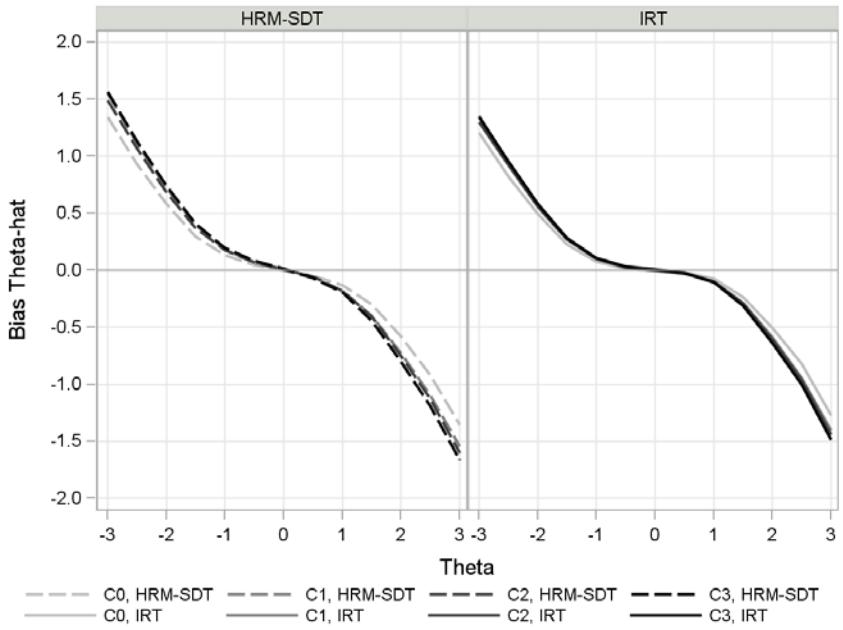

Figure 7. Conditional bias of proficiency estimates ($\hat{\theta}$) by condition and model, under posterior mode estimation across 30 replicates.

Figure 8 shows the conditional RMSE for all eight combinations of the four conditions and the two models. Two general patterns emerged with respect to RMSE: first, the RMSE of proficiency estimates tended to be similar across the four conditions for each model, with the exception being that Condition 0—under which all examinees responded to all items—tended to have the lowest RMSE. Secondly, the HRM-SDT estimates tended to have lower RMSE from about -1.5 to 1.5 on the known proficiency scale, while the naïve IRT model had lower RMSE outside that range. The greatest RMSE was observed for both the naïve IRT model and the HRM-SDT under Condition 3; specifically, from 0 to 2 on the known proficiency scale, the HRM-SDT RMSE ranged from 0.500 to 0.859 and the naïve IRT model RMSE ranged from 0.597 to 0.778.

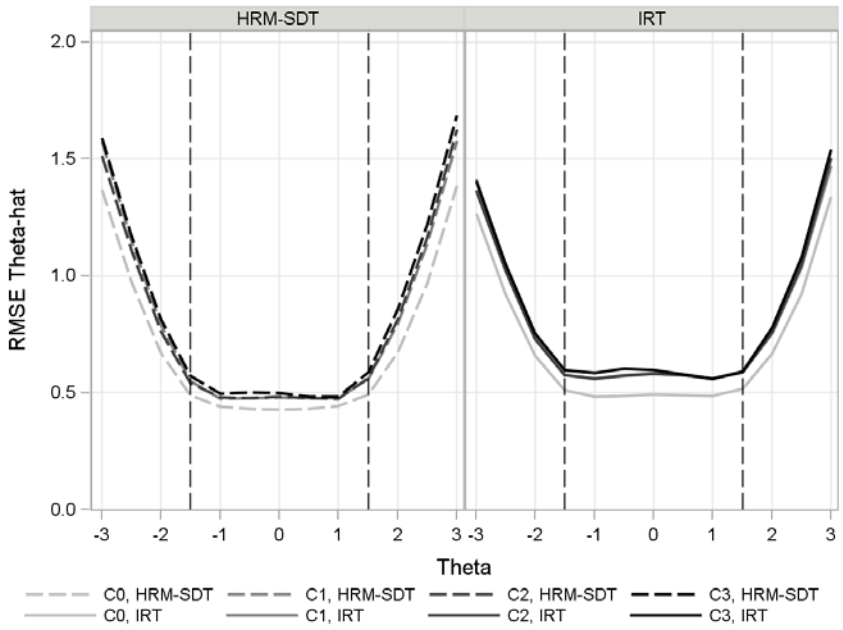

Figure 8. Conditional root-mean squared-error (RMSE) of proficiency estimates ($\hat{\theta}$) by condition and model, under posterior mode estimation across 30 replicates.

The consideration of differential recovery by selected item is just as important as considering the recovery of proficiency by the model used to estimate it. Figure 9 shows the empirical density of the deviation of estimated from true proficiency separately for examinees selecting Item 1 (the black, dotted density) and Item 2 (the gray, dashed density). Since examinees selected either Item 1 or 2 independent of proficiency, it is unsurprising that the deviations are centered at zero (i.e., the estimates do not exhibit systematic bias) and that the two densities have similar variance (i.e., the estimates do not show differential precision by selected item). These findings of near-zero bias are corroborated by Table 4. That table also shows the bias and RMSE of proficiency estimates by other key variables.

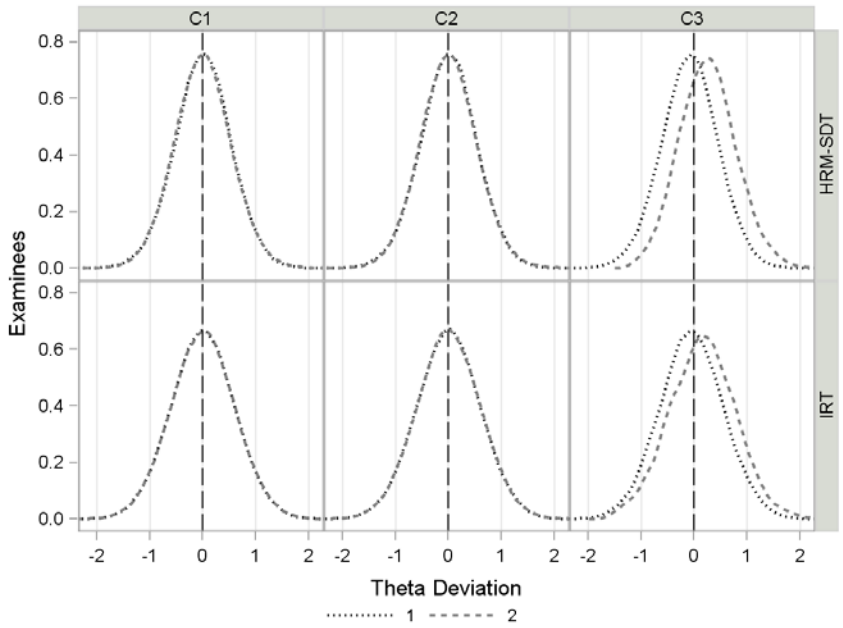

Figure 9. Deviation of estimates ($\hat{\theta}$) from true proficiency (θ) by model, condition, and selected item, under posterior mode estimation for 30 replicates.

Table 4.

Recovery of Examinee Proficiency by Selected Item, Model and Condition

				Bias		RMSE	
Cond.	**Group**	***M*(θ)**	***M*(*n*)**	**HRM**	**IRT**	**HRM**	**IRT**
0	Total	0.001	5000.0	-0.001	-0.002	0.485	0.524
1	Total	0.001	5000.0	-0.001	-0.003	0.544	0.607
	Selected Item 1	0.000	2497.5	-0.001	-0.006	0.543	0.606
	Selected Item 2	0.003	2502.5	-0.002	-0.001	0.545	0.608
2	Total	0.001	5000.0	-0.001	-0.002	0.543	0.606
	Selected Item 1	-0.001	3748.3	-0.001	-0.004	0.543	0.606
	Selected Item 2	0.008	1251.7	-0.002	0.002	0.544	0.606
	δ = 0	-0.001	2498.8	-0.003	-0.005	0.543	0.606
	δ = 1	0.004	2501.2	0.000	0.001	0.544	0.606
3	Total	0.001	5000.0	-0.001	-0.001	0.565	0.622
	Selected Item 1	0.267	3753.3	-0.088	-0.053	0.552	0.612
	Selected Item 2	-0.798	1246.7	0.260	0.155	0.604	0.649
	θ < 0	-0.796	2495.7	0.209	0.136	0.567	0.629
	θ ≥ 0	0.796	2504.3	-0.211	-0.138	0.564	0.615

Note. RMSE = root-mean squared error. Posterior mode estimation results for 30 replicates.

Graphically in Figure 9 and numerically in Table 4, the recovery of proficiency under the HRM-SDT is illustrated by selected item for Condition 2. The similarity between Conditions 1 and 2 is again seen in the small absolute magnitude of bias by selected item: no larger than 0.002, on average. The middle column of Figure 9 shows that, irrespective of item selected, estimated proficiency had similarly small bias and comparable precision. However, Condition 2 clearly diverges from Condition 1 in the relative rates of item selection. Where approximately equal numbers of examinees selected either Item 1 [$M(n) = 2{,}497.5$] or Item 2 [$M(n) = 2{,}502.5$] in Condition 1, about three times as many students selected Item 1 [$M(n) = 3{,}748.3$] as chose Item 2 [$M(n) = 1{,}251.7$] in Condition 2.

There are several ways in which the recovery of proficiency for examinees selecting different items in Conditions 1 and 2 diverges from the recovery for Condition 3. Table 4 shows that about 75% select the easier Item 1 and about 25% select the harder Item 2. However, where Conditions 1 and 2 had bias of negligible magnitudes for either those selecting Item 1 or Item 2, differential bias was observed for those groups in Condition 3. Figure 9's third column shows that there were clear differences in the expected bias of proficiency estimates by selected item. In particular, among those selecting Item 1 the mean bias of proficiency estimates was -0.088 and for those choosing Item 2 the mean bias of their proficiency estimates was 0.260. These larger values are due at least in part to the difference in mean proficiency for those groups—$M(\theta \mid \text{Selected Item 1}) = 0.267$ and $M(\theta \mid \text{Selected Item 2}) = -0.798$. Recall from Table 3 that under Condition 3, the mean bias for the proficiency estimates from the HRM-SDT around true proficiencies of -1.0 and 0.5—the nearest bins to the means of true proficiency for examinees selecting Items 2 and 1, respectively—were 0.196 and -0.071, respectively.

With the mechanism behind the examinees' selection of items being a central question, it was important to determine if differential bias or RMSE existed for proficiency estimates by the trait underlying selection. In other words, if the pattern of bias or RMSE differed for test-wise

(δ=1) and non-test-wise (δ=0) examinees in Condition 2 or for high-proficiency (θ>0) and low-proficiency (θ<0) students in Condition 3, then there may be additional concerns raised around allowing examinee-selected items. In Figure 10, the HRM-SDT estimate and true proficiency of examinees was plotted for one replicate and there appears little difference in the pattern for test-wise (δ=1) and non-test-wise (δ=0) examinees. Similarly, Table 4 shows little differential bias or RMSE, on average, by test-wisdom (δ) for Condition 2 across all 30 replicates. Also note that the difference in the relationship between estimated and true proficiency for high- (θ>0) and low-proficiency (θ<0) examinees in Condition 3 is basically symmetric about 0, as Figure 11 shows. In particular and across all 30 replicates, Table 4 shows that the estimated proficiency of examinees whose true proficiencies is near -1 tends to be overestimated [Bias($\hat{\theta}$) = 0.196], while those whose true proficiencies is above zero tends to be underestimated [Bias($\hat{\theta}$) = -0.198].

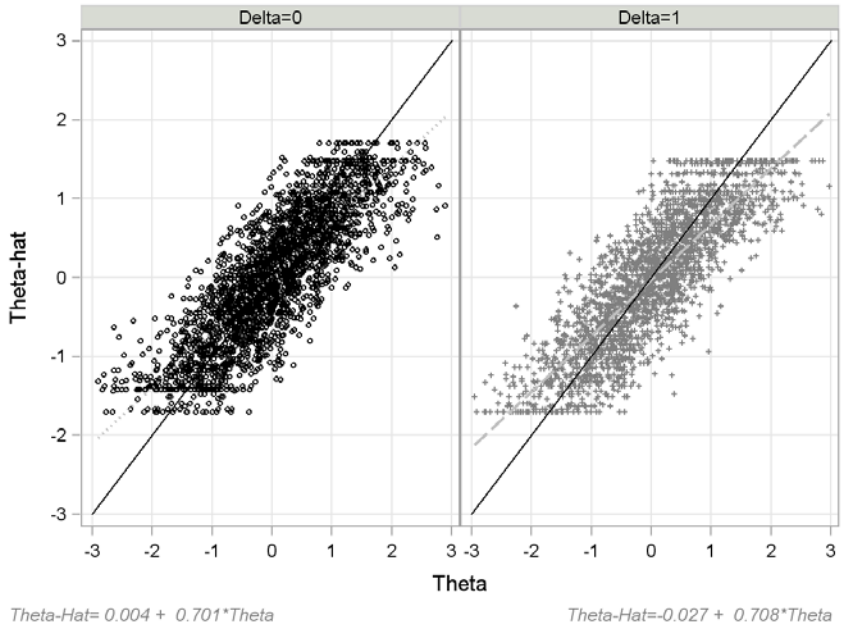

Figure 10. Proficiency estimates ($\hat{\theta}$) from the HRM-SDT by true proficiency (θ) and test wisdom (δ) for Condition 2 (test-wise item selection), under posterior mode estimation for replicate 1.

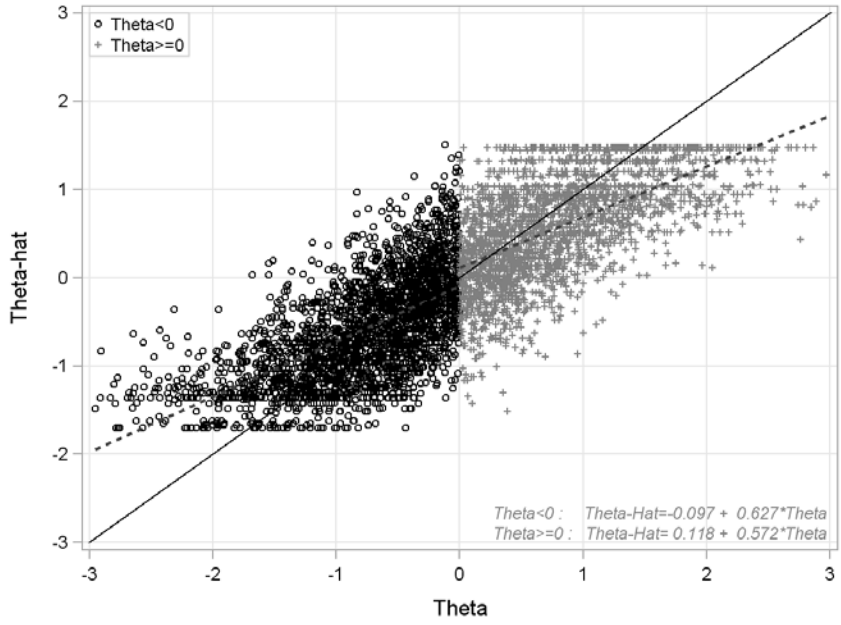

Figure 11. Proficiency estimates ($\hat{\theta}$) from the HRM-SDT by true proficiency (θ) and item selection threshold for θ for Condition 3, under posterior mode estimation for replicate 1.

Recovery of Latent Class Membership. Turning now to focus exclusively on the HRM-SDT, a feature that distinguishes this model from IRT models is the categorical latent variables that it employs. Weighted kappa and percent exact agreement are presented in Table 5 for the estimated and true latent class for the relevant item. The former statistic takes into account chance agreement, while the latter does not, but each provides similar information, with greater values indicating greater agreement. Table 5 shows some interesting patterns with respect to the recovery of latent class membership. As a comparison, under Condition 0 (in which all examinees answered all three items), weighted kappa for each item was about 0.892 and percent exact agreement was about 80.0%.

Table 5.

Classification Statistics by Item and Condition

Item	Variable	Weighted Kappa (Percent Exact Agreement)							
		Condition 0		Condition 1		Condition 2		Condition 3	
1	Estimated & True η_1	0.883	(79.0%)	0.648	(55.2%)	0.766	(66.9%)	0.724	(64.8%)
2	Estimated & True η_2	0.896	(81.1%)	0.653	(56.1%)	0.533	(44.2%)	0.318	(31.5%)
3	Estimated & True η_3	0.897	(79.8%)	0.895	(79.5%)	0.894	(79.4%)	0.893	(79.2%)

Note. Posterior mode estimation results for 30 replicates. Percent exact agreement is shown in parentheses. Estimated η_i were most likely classes. Calculations for kappa used Cicchetti-Allison weights (Cicchetti & Allison, 1971).

For Item 3, the only item required of all simulated examinees, there was minimal variability across the four conditions: kappa only varied from 0.893 to 0.897 and the percent exact agreement only varied from 79.2% to 79.8%. In Condition 1 the kappas (percent exact agreement) were fairly consistent for Items 1 and 2—the two items subject to examinee selection—at 0.648 to 0.653 (55.2% to 56.1%), respectively. It was in Conditions 2 and 3 where differences emerged in agreement for the two items subject to examinee selection. In Condition 2, kappa (percent exact agreement) for Item 1 was 0.766 (66.9%) and for Item 2 it was 0.533 (44.2%). An even greater difference was observed for Condition 3, where kappa (percent exact agreement) for Item 1 was 0.724 (64.8%) and for Item 2 it was 0.318 (31.5%). This shows that latent class membership for the items under examinee selection were recovered relatively less well than those of the required item. And it seems that the differences between examinee-selected and required items were more extreme in the two conditions where the response rates to examinee-selected items differed (i.e., Conditions 2 and 3).

4.2. Item Parameter Recovery

Before considering the recovery of item parameters when examinees selected according to three possible mechanisms, some comments on the recovery of item parameters under the full data condition (i.e., Condition 0) are warranted. Table 6 shows that there was negligible bias in the estimates of the three item discrimination parameters (a_i).

Table 6.

Recovery of Item Parameters for Condition 0

	True	Estimate			Standard Error			
Par.	Value	*M*	Bias	|% B|	SE_{Pop}	*M*	Bias	|% B|
a_1	1.0	1.004	0.004	0.421	0.060	0.053	-0.007	12.250
a_2	1.0	1.020	0.020	1.986	0.047	0.055	0.008	17.136
a_3	1.5	1.479	-0.021	1.385	0.098	0.097	-0.001	0.634
b_{11}	-1.7	-1.620	0.080	4.692	0.201	0.264	0.063	31.352
b_{12}	-1.1	-1.100	0.000	0.043	0.122	0.134	0.013	10.465
b_{13}	-0.5	-0.515	-0.015	2.931	0.096	0.115	0.018	18.869
b_{14}	0.1	0.114	0.014	13.573	0.096	0.101	0.005	5.746
b_{15}	0.7	0.700	0.000	0.025	0.156	0.156	0.000	0.152
b_{21}	-0.7	-0.749	-0.049	7.040	0.182	0.178	-0.003	1.865
b_{22}	-0.1	-0.115	-0.015	14.687	0.126	0.114	-0.013	10.063
b_{23}	0.5	0.484	-0.016	3.121	0.139	0.126	-0.013	9.099
b_{24}	1.1	1.129	0.029	2.594	0.197	0.151	-0.045	23.069
b_{25}	1.7	1.624	-0.076	4.467	0.200	0.258	0.058	29.215
b_{31}	-1.6	-1.521	0.079	4.933	0.200	0.214	0.014	7.214
b_{32}	-0.8	-0.792	0.008	1.049	0.141	0.127	-0.014	10.243
b_{33}	0.0	-0.017	-0.017	n/a	0.123	0.118	-0.004	3.544
b_{34}	0.8	0.824	0.024	2.993	0.137	0.128	-0.009	6.347
b_{35}	1.6	1.548	-0.052	3.266	0.226	0.226	0.000	0.029

Note. |% B| = absolute percent bias; SE_{Pop} is the *SD* of the estimates. Posterior mode estimation results for 30 replicates.

Additional relevant information appears in the the top-left panel of Figure 12, where the empirical densities of deviations of $\hat{a}_i$ from the true values are plotted. With each of the three densities centered at zero and the relatively small variance in these deviations, the good recovery of item discrimination is apparent. Almost as good was the recovery of the 15 item

location parameters (b_{ik}), whose estimates only exhibited large bias twice and moderate bias once. It is also worth noting that these poorly recovered parameters occurred where the true parameter values were quite small in absolute value (either 0.1 or 0.7,) so that any small bias appeared larger in terms of the absolute percent bias. Note that for the true item location equal to 0 (i.e., b_{33}) the absolute percent bias cannot be computed, so the absolute percent bias of the remaining 14 location parameter estimates are discussed herein. The recovery of this parameter may however be inspected graphically in the top row of Figure 13, where the deviation of estimated from true b_{ik} is shown. It shows that the estimates of b_{ik} generally do not systematically vary from the true values. The standard errors of parameter estimates in Condition 0 were generally over-estimated relative to population standard error (SE_{Pop}; i.e., standard deviation of parameter estimates). In particular, the standard errors of two of three discrimination parameters and 7 of 14 location parameters had large absolute percent bias, with another four standard errors of item locations showing moderate bias.

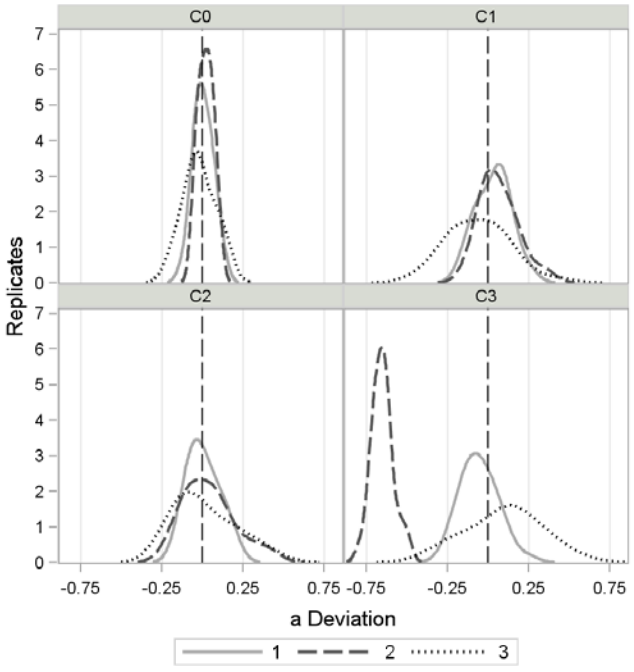

Figure 12. Density of estimate deviation from true item discrimination (a_i) by condition and item.

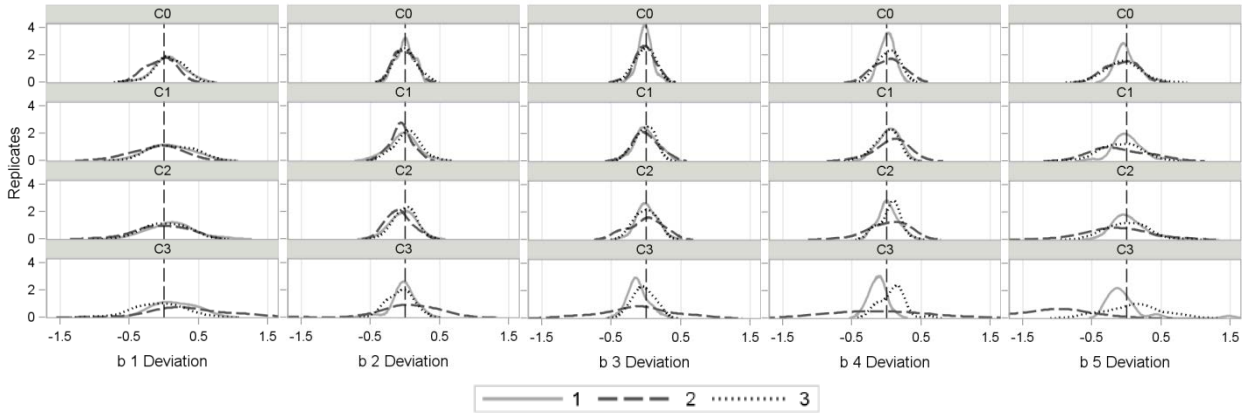

Figure 13. Density of estimate deviation from true item location (b_{ik}) by condition and item.

Table 7 shows the recovery of HRM-SDT item parameters for Condition 1, in which examinees were simulated to have randomly and with equal expected frequency chosen between Items 1 and 2. There was some moderate bias in one discrimination parameter (|% Bias{$\hat{a}_2$}| = 6.484%), but the other two item discrimination parameter estimates had relatively little bias. In the top-right panel of Figure 12, where with the three densities roughly centered at zero, the moderate bias of $\hat{a}_2$ is put into perspective. In terms of item location parameter estimates, 4 of 14 had moderate absolute percent bias—ranging from 5.631% to 7.288%—and three had large absolute percent bias—ranging from 15.149% to 59.377%. It should be noted that the two most extreme locations in terms of absolute percent bias ($\hat{b}_{14}$ and $\hat{b}_{22}$) were not all that large in absolute terms—with estimated bias of 0.034 and -0.059, respectively. The reason the absolute percent bias is so large for these parameters is that the true parameters themselves are quite small, 0.1 and -0.1 respectively, and this led to large percent bias even with relatively small bias. A graphical representation of the recovery of b_{ik} is given in the second row of Figure 13 for Condition 1; note that the estimates of b_{ik} generally do not systematically vary from the true values. The standard errors of all discrimination parameters and all but four location parameters were overestimated, with large absolute percent bias.

Table 7.

Recovery of Item Parameters for Condition 1

	True	Estimate			Standard Error			
Par.	Value	*M*	Bias	|% B|	SE_{Pop}	*M*	Bias	|% B|
a_1	1.0	1.032	0.032	3.213	0.102	0.184	0.082	80.358
a_2	1.0	1.065	0.065	6.484	0.118	0.195	0.077	65.325
a_3	1.5	1.438	-0.062	4.161	0.186	0.344	0.158	84.992
b_{11}	-1.7	-1.617	0.083	4.902	0.289	0.417	0.127	44.004
b_{12}	-1.1	-1.134	-0.034	3.096	0.193	0.250	0.057	29.249
b_{13}	-0.5	-0.493	0.007	1.351	0.154	0.173	0.019	12.347
b_{14}	0.1	0.134	0.034	34.153	0.147	0.154	0.007	4.960
b_{15}	0.7	0.734	0.034	4.804	0.231	0.271	0.040	17.500
b_{21}	-0.7	-0.806	-0.106	15.149	0.314	0.300	-0.013	4.300
b_{22}	-0.1	-0.159	-0.059	59.377	0.148	0.169	0.021	14.478
b_{23}	0.5	0.498	-0.002	0.443	0.175	0.187	0.013	7.162
b_{24}	1.1	1.171	0.071	6.415	0.235	0.272	0.037	15.947
b_{25}	1.7	1.604	-0.096	5.631	0.345	0.422	0.077	22.219
b_{31}	-1.6	-1.483	0.117	7.288	0.274	0.440	0.166	60.450
b_{32}	-0.8	-0.767	0.033	4.186	0.173	0.208	0.035	20.265
b_{33}	0.0	-0.021	-0.021	n/a	0.134	0.126	-0.008	6.252
b_{34}	0.8	0.796	-0.004	0.480	0.159	0.212	0.053	33.178
b_{35}	1.6	1.506	-0.094	5.880	0.273	0.443	0.171	62.537

Note. |% B| = absolute percent bias; SE_{Pop} is the *SD* of the estimates. Posterior mode estimation results for 30 replicates.

The recovery of item parameters for Condition 2 is shown in Table 8. Given its similarity to Condition 1—in that examinees selected either Item 1 or 2 independent of their proficiency—a certain amount of similarity was expected. Specifically, none of the three item discrimination estimates tended to exhibit even moderate bias—also reinforced by the bottom-left panel of Figure 12—and three item thresholds exhibited moderate absolute percent bias (ranging from 5.672% to 8.765%) and two exhibited large absolute percent bias, with values of 20.637% and 92.057%. Those two locations exhibiting large absolute percent bias were again associated with the location parameters that were smallest in absolute value (b_{14} and b_{22}); the biases were 0.021 and -0.092, respectively. These two item location parameters also stand out when

considering the third row of Figure 13. In terms of the estimated variance of the parameter estimates, all discrimination parameters and all but two location parameters exhibited at least moderate bias, again with standard errors having been over-estimated, relative to the population standard error.

Table 8.

Recovery of Item Parameters for Condition 2

	True	Estimate			Standard Error			
Par.	Value	*M*	Bias	|% B|	SE_{Pop}	*M*	Bias	|% B|
a_1	1.0	1.005	0.005	0.521	0.097	0.167	0.070	72.610
a_2	1.0	1.039	0.039	3.943	0.161	0.198	0.037	23.137
a_3	1.5	1.516	0.016	1.092	0.185	0.372	0.187	101.511
b_{11}	-1.7	-1.592	0.108	6.339	0.309	0.357	0.048	15.569
b_{12}	-1.1	-1.128	-0.028	2.539	0.163	0.213	0.050	30.792
b_{13}	-0.5	-0.486	0.014	2.714	0.140	0.142	0.002	1.681
b_{14}	0.1	0.121	0.021	20.637	0.142	0.122	-0.020	13.966
b_{15}	0.7	0.730	0.030	4.251	0.225	0.235	0.010	4.557
b_{21}	-0.7	-0.729	-0.029	4.163	0.345	0.405	0.060	17.320
b_{22}	-0.1	-0.192	-0.092	92.057	0.168	0.237	0.069	40.946
b_{23}	0.5	0.487	-0.013	2.559	0.233	0.269	0.035	15.205
b_{24}	1.1	1.119	0.019	1.737	0.265	0.343	0.078	29.571
b_{25}	1.7	1.551	-0.149	8.765	0.434	0.510	0.077	17.637
b_{31}	-1.6	-1.576	0.024	1.512	0.272	0.458	0.187	68.667
b_{32}	-0.8	-0.807	-0.007	0.857	0.148	0.215	0.067	45.386
b_{33}	0.0	-0.008	-0.008	n/a	0.136	0.124	-0.012	8.510
b_{34}	0.8	0.845	0.045	5.672	0.149	0.223	0.073	49.260
b_{35}	1.6	1.609	0.009	0.570	0.284	0.469	0.185	65.249

Note. |% B| = absolute percent bias; SE_{Pop} is the *SD* of the estimates. Posterior mode estimation results for 30 replicates.

Turning now to Condition 3, the item discrimination and location parameters are not recovered as well as in the two earlier conditions. Table 9 shows that one discrimination parameter exhibited large absolute percent bias (65.254% for $\hat{a}_2$) and the other two had moderate absolute percent bias (5.646% for $\hat{a}_1$ and 6.340% for $\hat{a}_3$). The same a_i estimates,

presented in Figure 12's bottom-right panel, show extreme bias. Eleven of fourteen item location parameters had at least moderate percent bias (≥ 6.110%), with 7 of those being large (≥ 10.217%). It is interesting to note that Item 2—the harder of the two items subject to selection—has the poorest recovery; perhaps that is due to the fact that fewer examinees were expected to answer it under this condition. In other words, this item has the most missing data and that missingness was simulated to be directly related to examinee proficiency, so this poor recovery is to be expected. This sparseness of data is also probably the cause of the great variance in the estimates of b_{2k} around the true value, shown in the bottom row of Figure 13. The difference in underlying proficiency among examinees selecting Items 1 and 2 likely causes the bias for b_{1k} and b_{2k}, apparent in that same plot. As in the other two conditions, standard errors for item parameter estimates in Condition 3 tended to be over-estimated relative to the population standard error, with two of three standard errors for the discrimination parameters showing large absolute percent bias and 13 of the 14 location parameter standard errors showing at least moderate absolute percent bias (≥ 6.281%).

Table 9.

Recovery of Item Parameters for Condition 3

	True	Estimate			Standard Error			
Par.	Value	*M*	Bias	|% B|	SE_{Pop}	*M*	Bias	|% B|
a_1	1.0	0.944	-0.056	5.646	0.110	0.157	0.047	42.784
a_2	1.0	0.347	-0.653	65.254	0.066	0.070	0.003	4.904
a_3	1.5	1.595	0.095	6.340	0.224	0.444	0.220	98.054
b_{11}	-1.7	-1.539	0.161	9.454	0.270	0.378	0.108	40.162
b_{12}	-1.1	-1.124	-0.024	2.185	0.181	0.230	0.048	26.731
b_{13}	-0.5	-0.619	-0.119	23.789	0.145	0.154	0.009	6.518
b_{14}	0.1	-0.029	-0.129	129.020	0.121	0.119	-0.002	1.889
b_{15}	0.7	0.668	-0.032	4.627	0.340	0.223	-0.117	34.444
b_{21}	-0.7	-0.274	0.426	60.856	0.604	0.986	0.382	63.366
b_{22}	-0.1	-0.163	-0.063	63.453	0.685	0.534	-0.151	22.051
b_{23}	0.5	0.382	-0.118	23.575	0.485	0.449	-0.036	7.391
b_{24}	1.1	1.015	-0.085	7.760	0.674	0.684	0.011	1.589
b_{25}	1.7	0.492	-1.208	71.051	0.655	0.881	0.226	34.498
b_{31}	-1.6	-1.672	-0.072	4.471	0.337	0.544	0.207	61.543
b_{32}	-0.8	-0.869	-0.069	8.584	0.168	0.249	0.081	47.855
b_{33}	0.0	-0.035	-0.035	n/a	0.144	0.129	-0.015	10.510
b_{34}	0.8	0.882	0.082	10.217	0.174	0.258	0.084	48.051
b_{35}	1.6	1.698	0.098	6.110	0.393	0.544	0.152	38.667

Note. |% B| = absolute percent bias; SE_{Pop} is the *SD* of the estimates. Posterior mode estimation results for 30 replicates.

4.3. Rater Parameter Recovery

Another aspect of the HRM-SDT that differentiates it from the typical IRT model is its explicit rater components. Table 10 shows the recovery of rater parameters for Condition 0, which is helpful when considering the conditions in which examinees select from among Items 1 and 2. All rater discrimination (d_{ij}) parameters were recovered with negligible absolute percent bias. Separating the lower- and higher-discrimination raters into two columns, the deviation of estimated from true d_{ij} is plotted in Figure 14. The d_{ij} estimates for both lower- and higher-discrimination raters are shown in the top row to be recovered accurately. Only three rater criteria estimates ($\hat{c}_{ijk}$) exhibited even moderate absolute percent bias (≤ 8.529%) and the distribution of estimated c_{ijk} may be examined more closely in the top rows of each sub-figure in Figure 15. The estimated standard errors of the rater parameters on the other hand were overestimated quite often. The bias for the standard error of 3 rater discrimination parameter estimates was large in terms of absolute percent bias and it was moderate for one. Of the 30 rater criteria, the standard error of 25 exhibited at least moderate absolute percent bias.

Table 10.

Recovery of Rater Parameters for Condition 0

	True	Estimate			Standard Error			
Par.	Value	*M*	Bias	|% B|	SE_{Pop}	*M*	Bias	|% B|
d_{11}	2.2	2.196	-0.004	0.175	0.073	0.086	0.012	16.872
d_{12}	3.8	3.738	-0.062	1.624	0.172	0.232	0.060	34.834
d_{23}	1.8	1.829	0.029	1.622	0.051	0.067	0.016	30.682
d_{24}	4.2	4.023	-0.177	4.217	0.283	0.295	0.011	4.054
d_{35}	2.0	2.012	0.012	0.596	0.065	0.067	0.002	3.546
d_{36}	4.0	3.955	-0.045	1.129	0.276	0.252	-0.024	8.737
c_{111}	1.1	1.039	-0.061	5.517	0.176	0.230	0.054	30.583
c_{112}	3.3	3.256	-0.044	1.336	0.213	0.251	0.038	17.792
c_{113}	5.5	5.464	-0.036	0.653	0.218	0.285	0.067	30.884
c_{114}	7.7	7.679	-0.021	0.269	0.265	0.326	0.061	22.956
c_{115}	9.9	9.870	-0.030	0.304	0.299	0.368	0.069	23.058
c_{121}	1.9	1.738	-0.162	8.529	0.379	0.462	0.083	21.946
c_{122}	5.7	5.536	-0.164	2.883	0.404	0.531	0.127	31.361
c_{123}	9.5	9.289	-0.211	2.222	0.528	0.693	0.165	31.318
c_{124}	13.3	13.030	-0.270	2.031	0.657	0.877	0.220	33.398
c_{125}	17.1	16.807	-0.293	1.711	0.756	1.064	0.309	40.830
c_{231}	0.9	0.938	0.038	4.239	0.105	0.126	0.022	20.835
c_{232}	2.7	2.774	0.074	2.743	0.130	0.149	0.019	14.886
c_{233}	4.5	4.614	0.114	2.527	0.143	0.178	0.035	24.608
c_{234}	6.3	6.440	0.140	2.228	0.188	0.210	0.022	11.859
c_{235}	8.1	8.285	0.185	2.282	0.219	0.246	0.027	12.271
c_{241}	2.1	2.107	0.007	0.310	0.381	0.351	-0.030	7.807
c_{242}	6.3	6.128	-0.172	2.726	0.420	0.503	0.083	19.863
c_{243}	10.5	10.178	-0.322	3.069	0.695	0.759	0.064	9.259
c_{244}	14.7	14.241	-0.459	3.125	0.964	1.034	0.069	7.197
c_{245}	18.9	18.357	-0.543	2.875	1.210	1.319	0.108	8.934
c_{351}	1.0	0.977	-0.023	2.271	0.137	0.136	-0.001	0.866
c_{352}	3.0	2.994	-0.006	0.215	0.168	0.162	-0.007	3.915
c_{353}	5.0	5.005	0.005	0.107	0.223	0.195	-0.028	12.528
c_{354}	7.0	7.032	0.032	0.450	0.248	0.232	-0.016	6.281
c_{355}	9.0	9.048	0.048	0.535	0.291	0.271	-0.020	6.709
c_{361}	2.0	1.847	-0.153	7.632	0.266	0.315	0.049	18.236
c_{362}	6.0	5.880	-0.120	2.003	0.442	0.453	0.011	2.541
c_{363}	10.0	9.841	-0.159	1.592	0.590	0.664	0.074	12.501
c_{364}	14.0	13.854	-0.146	1.041	0.902	0.896	-0.006	0.672
c_{365}	18.0	17.830	-0.170	0.942	1.128	1.132	0.004	0.350

Note. |% B| = absolute percent bias; SE_{Pop} is the *SD* of the estimates. Posterior mode estimation results for 30 replicates.

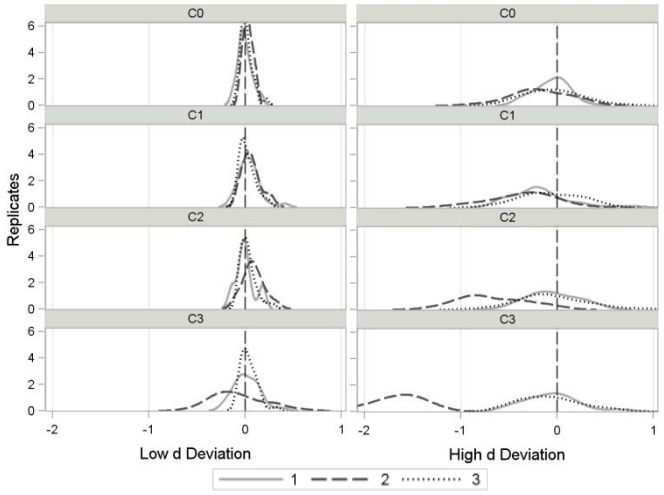

Figure 14. Density of estimate deviation from true rater discrimination (d_{ij}) by condition and item.

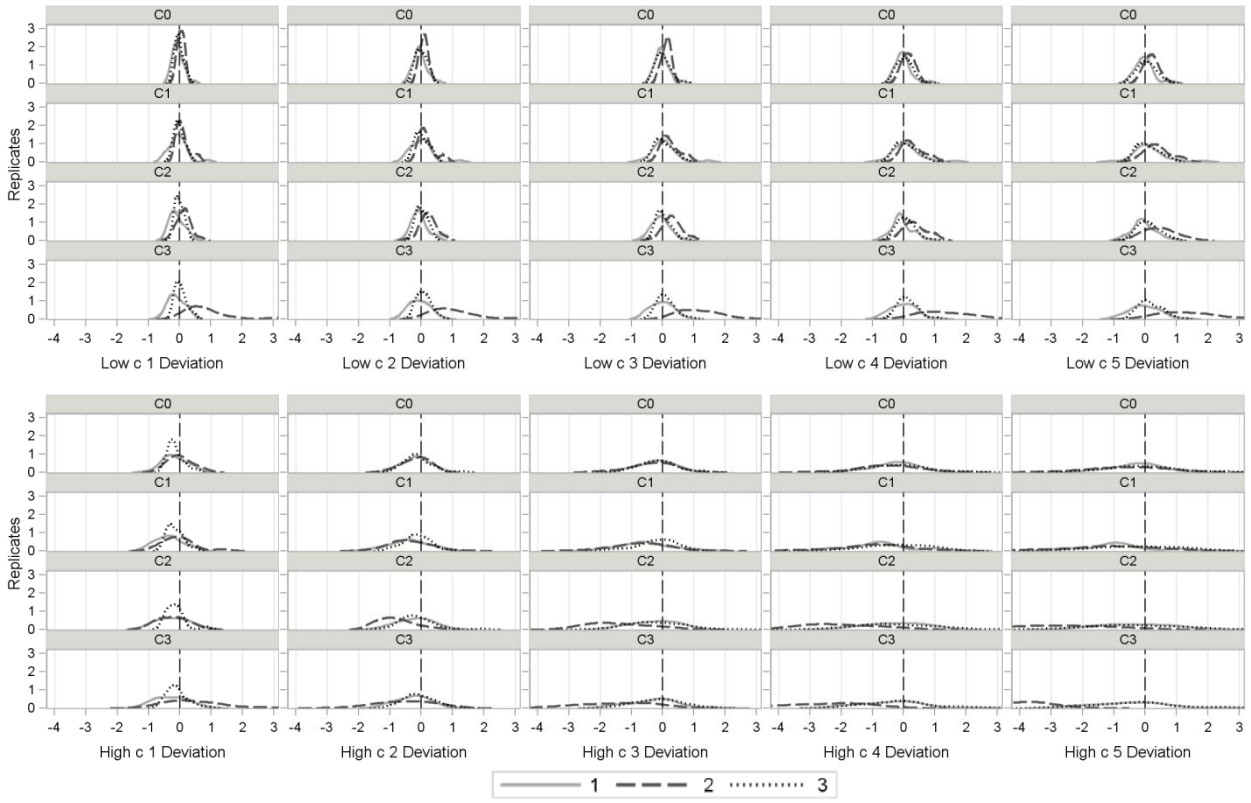

Figure 15. Density of estimate deviation from true rater criteria (c_{ijk}) by condition and item.

Most relevant to this study is the consideration of the three conditions in which examinees selected from the pair of possible items. Table 11 shows the recovery of rater parameters for Condition 1. One rater discrimination parameter estimate exhibited moderate bias (|% Bias{$\hat{d}_{24}$}| = 8.669%) and the remaining had negligible absolute percent bias. This pattern is visible in the second row of Figure 14. Only two rater criteria estimates exhibited large absolute percent bias ($\hat{c}_{121}$ at 18.525% and $\hat{c}_{231}$ at 10.436%) and 11 of the remaining 28 rater criteria exhibited moderate absolute percent bias (≤ 8.010%). This generally good recovery of criteria is supported by a closer examination of the second rows of each sub-figure to Figure 15. The rater discrimination and criteria parameters for the pair of raters rating Item 3—the required item—were recovered relatively better than for either pair rating Item 1 or Item 2—those subject to examinee-item selection. The bias in the estimated standard error for 2 of 6 rater discrimination parameters were large in terms of absolute percent bias, while 23 of the 30 rater criteria exhibited at least moderate absolute percent bias for the estimated standard error.

Table 11.

Recovery of Rater Parameters for Condition 1

	True	Estimate			Standard Error			
Par.	Value	*M*	Bias	|% B|	SE_{Pop}	*M*	Bias	|% B|
d_{11}	2.2	2.240	0.040	1.801	0.112	0.129	0.017	14.883
d_{12}	3.8	3.635	-0.165	4.332	0.325	0.326	0.002	0.572
d_{23}	1.8	1.864	0.064	3.540	0.096	0.095	-0.001	1.212
d_{24}	4.2	3.836	-0.364	8.669	0.327	0.382	0.055	16.740
d_{35}	2.0	2.021	0.021	1.070	0.082	0.080	-0.002	2.619
d_{36}	4.0	3.930	-0.070	1.744	0.294	0.303	0.009	2.994
c_{111}	1.1	1.042	-0.058	5.253	0.285	0.302	0.017	5.992
c_{112}	3.3	3.284	-0.016	0.478	0.342	0.336	-0.006	1.756
c_{113}	5.5	5.552	0.052	0.947	0.390	0.387	-0.003	0.693
c_{114}	7.7	7.794	0.094	1.215	0.461	0.448	-0.013	2.805
c_{115}	9.9	10.027	0.127	1.282	0.527	0.512	-0.015	2.836
c_{121}	1.9	1.548	-0.352	18.525	0.424	0.564	0.140	33.053
c_{122}	5.7	5.308	-0.392	6.869	0.594	0.693	0.099	16.718
c_{123}	9.5	8.942	-0.558	5.874	0.814	0.895	0.081	9.990
c_{124}	13.3	12.595	-0.705	5.302	1.073	1.134	0.061	5.698
c_{125}	17.1	16.299	-0.801	4.686	1.257	1.386	0.129	10.269
c_{231}	0.9	0.994	0.094	10.436	0.217	0.187	-0.030	13.806
c_{232}	2.7	2.860	0.160	5.910	0.256	0.224	-0.032	12.437
c_{233}	4.5	4.741	0.241	5.345	0.291	0.269	-0.022	7.568
c_{234}	6.3	6.594	0.294	4.662	0.339	0.318	-0.021	6.261
c_{235}	8.1	8.484	0.384	4.736	0.401	0.372	-0.029	7.164
c_{241}	2.1	2.057	-0.043	2.064	0.530	0.451	-0.079	14.877
c_{242}	6.3	5.934	-0.366	5.809	0.654	0.679	0.025	3.848
c_{243}	10.5	9.803	-0.697	6.640	0.916	0.995	0.079	8.621
c_{244}	14.7	13.697	-1.003	6.824	1.206	1.354	0.148	12.240
c_{245}	18.9	17.665	-1.235	6.537	1.473	1.733	0.261	17.696
c_{351}	1.0	0.991	-0.009	0.869	0.163	0.163	0.000	0.188
c_{352}	3.0	3.014	0.014	0.476	0.202	0.194	-0.009	4.286
c_{353}	5.0	5.035	0.035	0.694	0.268	0.233	-0.034	12.865
c_{354}	7.0	7.068	0.068	0.975	0.304	0.278	-0.026	8.620
c_{355}	9.0	9.089	0.089	0.990	0.352	0.324	-0.028	7.910
c_{361}	2.0	1.840	-0.160	8.010	0.253	0.369	0.116	45.611
c_{362}	6.0	5.845	-0.155	2.588	0.422	0.540	0.119	28.180
c_{363}	10.0	9.793	-0.207	2.074	0.622	0.798	0.175	28.143
c_{364}	14.0	13.786	-0.214	1.528	0.984	1.080	0.096	9.771
c_{365}	18.0	17.728	-0.272	1.514	1.285	1.363	0.078	6.041

Note. |% B| = absolute percent bias; SE_{Pop} is the *SD* of the estimates. Posterior mode estimation results for 30 replicates.

The similarity of Table 12 and Table 11 reflect the inherent similarity of Conditions 2 and 1. In particular, under Condition 2, the rater discrimination parameters were recovered with negligible absolute percent bias for all but one rater (4), whose bias was large (|% Bias$\{\hat{d}_{24}\}$| = 15.891%) and was problematic in Condition 1 as well. The clear negative bias of that parameter estimate also appeared in the third row of Figure 14. Only 3 of the 20 rater criteria for raters rating Items 1 and 3 exhibited even moderate absolute percent bias, but all 10 rater criteria for raters rating Item 2 were at least moderate in terms of absolute percent bias. Note the positive shift of the densities associated with Item 2 and Rater 3—the medium-gray dashed densities in the third row of the first sub-figure to Figure 14. The negative bias of the other rater for Item 2 is shown in the third row of the second sub-figure. The estimated standard errors were also recovered poorly for Condition 2, as they were for Condition 1. The estimated standard errors of the rater discrimination parameters were large for 4 raters and moderate for 1 rater. In terms of the estimated standard errors of rater criteria, 21 were recovered with large absolute percent bias and another 7 had moderate bias.

Table 12.

Recovery of Rater Parameters for Condition 2

	True	Estimate			Standard Error			
Par.	Value	*M*	Bias	|% B|	SE_{Pop}	*M*	Bias	|% B|
d_{11}	2.2	2.200	0.000	0.021	0.100	0.103	0.003	3.437
d_{12}	3.8	3.755	-0.045	1.191	0.244	0.291	0.048	19.529
d_{23}	1.8	1.885	0.085	4.724	0.108	0.136	0.027	25.369
d_{24}	4.2	3.533	-0.667	15.891	0.344	0.455	0.112	32.466
d_{35}	2.0	2.028	0.028	1.382	0.084	0.078	-0.006	6.869
d_{36}	4.0	3.939	-0.061	1.532	0.349	0.299	-0.050	14.267
c_{111}	1.1	1.011	-0.089	8.112	0.225	0.246	0.021	9.411
c_{112}	3.3	3.233	-0.067	2.035	0.252	0.274	0.022	8.618
c_{113}	5.5	5.436	-0.064	1.160	0.281	0.317	0.036	12.815
c_{114}	7.7	7.649	-0.051	0.656	0.331	0.367	0.036	10.913
c_{115}	9.9	9.854	-0.046	0.469	0.374	0.420	0.047	12.497
c_{121}	1.9	1.666	-0.234	12.317	0.478	0.488	0.010	2.170
c_{122}	5.7	5.506	-0.194	3.400	0.537	0.613	0.076	14.198
c_{123}	9.5	9.242	-0.258	2.720	0.675	0.811	0.136	20.102
c_{124}	13.3	13.007	-0.293	2.205	0.878	1.039	0.160	18.257
c_{125}	17.1	16.787	-0.313	1.831	1.035	1.274	0.239	23.113
c_{231}	0.9	1.031	0.131	14.589	0.228	0.286	0.058	25.450
c_{232}	2.7	2.917	0.217	8.036	0.261	0.332	0.071	27.338
c_{233}	4.5	4.801	0.301	6.679	0.304	0.390	0.085	27.919
c_{234}	6.3	6.668	0.368	5.842	0.374	0.454	0.079	21.185
c_{235}	8.1	8.562	0.462	5.700	0.480	0.526	0.046	9.506
c_{241}	2.1	1.859	-0.241	11.455	0.454	0.646	0.192	42.233
c_{242}	6.3	5.487	-0.813	12.904	0.573	0.865	0.292	50.939
c_{243}	10.5	9.083	-1.417	13.496	0.915	1.207	0.292	31.913
c_{244}	14.7	12.675	-2.025	13.774	1.089	1.581	0.492	45.175
c_{245}	18.9	16.269	-2.631	13.918	1.340	1.971	0.631	47.082
c_{351}	1.0	0.992	-0.008	0.764	0.135	0.151	0.016	11.838
c_{352}	3.0	3.015	0.015	0.514	0.177	0.182	0.005	2.627
c_{353}	5.0	5.033	0.033	0.667	0.242	0.222	-0.020	8.385
c_{354}	7.0	7.070	0.070	0.994	0.282	0.267	-0.015	5.308
c_{355}	9.0	9.097	0.097	1.077	0.333	0.314	-0.019	5.833
c_{361}	2.0	1.864	-0.136	6.813	0.329	0.353	0.023	7.091
c_{362}	6.0	5.853	-0.147	2.448	0.584	0.519	-0.066	11.214
c_{363}	10.0	9.786	-0.214	2.138	0.871	0.776	-0.096	10.971
c_{364}	14.0	13.775	-0.225	1.605	1.247	1.055	-0.192	15.422
c_{365}	18.0	17.725	-0.275	1.527	1.539	1.340	-0.199	12.929

Note. |% B| = absolute percent bias; SE_{Pop} is the *SD* of the estimates. Posterior mode estimation results for 30 replicates.

Finally, the recovery of rater parameters in Condition 3 is quite similar to that of Condition 2. Table 13 shows the recovery of rater parameters for Condition 3. As in the previous two conditions, all but one rater discrimination parameter was recovered with negligible absolute percent bias; the exception was once again Rater 4, whose estimated discrimination ($\hat{d}_{24}$) had 40.210% absolute bias. This severe bias is clear in the fourth row and second column of Figure 14. In terms of rater criteria, again, for the 4 raters rating either Item 1 or Item 3, only two criteria were recovered with moderate absolute percent bias and only one with large absolute percent bias. It is for Item 2 where 9 of 10 criteria were recovered with large absolute percent bias. These biases are clear in the fourth rows of both sub-figures to Figure 15; in other words, the criteria for the lower-discrimination rater of Item 2 were positively biased, while those of the higher-discrimination rater of that same item tended to be negatively biased. The poor recovery of the standard errors of rater parameters was also evident in this table. The standard error of three raters' discriminations had large absolute percent bias. The standard errors of raters' criteria were estimated with large absolute percent bias for 14 criteria and moderate for 6 criteria. Where rater parameter recovery is quite similar for Condition 3, relative to the other two conditions, the same cannot be said of item parameter recovery, which was substantially worse in Condition 3, relative to the others. This is likely due to the fact that the item parameter recovery is made worse when examinee-item selection is related to proficiency—as it is in Condition 3—but this does not cause similar problems in the signal detection theory (i.e., rater) portion of the model.

Table 13.

Recovery of Rater Parameters for Condition 3

	True	Estimate			Standard Error			
Par.	Value	*M*	Bias	|% B|	SE_{Pop}	*M*	Bias	|% B|
d_{11}	2.2	2.215	0.015	0.686	0.133	0.108	-0.025	18.966
d_{12}	3.8	3.745	-0.055	1.454	0.282	0.296	0.014	4.865
d_{23}	1.8	1.724	-0.076	4.211	0.273	0.267	-0.006	2.137
d_{24}	4.2	2.511	-1.689	40.210	0.306	0.495	0.189	61.561
d_{35}	2.0	2.035	0.035	1.766	0.083	0.087	0.004	4.478
d_{36}	4.0	3.921	-0.079	1.963	0.300	0.332	0.033	10.972
c_{111}	1.1	0.997	-0.103	9.342	0.245	0.274	0.029	11.727
c_{112}	3.3	3.239	-0.061	1.853	0.292	0.304	0.012	4.162
c_{113}	5.5	5.470	-0.030	0.542	0.348	0.349	0.001	0.212
c_{114}	7.7	7.694	-0.006	0.072	0.407	0.401	-0.006	1.476
c_{115}	9.9	9.887	-0.013	0.132	0.477	0.453	-0.025	5.190
c_{121}	1.9	1.627	-0.273	14.383	0.482	0.543	0.061	12.668
c_{122}	5.7	5.427	-0.273	4.788	0.517	0.646	0.129	24.923
c_{123}	9.5	9.223	-0.277	2.914	0.813	0.857	0.044	5.459
c_{124}	13.3	12.984	-0.316	2.379	1.082	1.089	0.006	0.578
c_{125}	17.1	16.722	-0.378	2.208	1.308	1.308	0.001	0.047
c_{231}	0.9	1.750	0.850	94.436	0.757	0.715	-0.043	5.622
c_{232}	2.7	3.783	1.083	40.115	0.825	0.762	-0.063	7.659
c_{233}	4.5	5.813	1.313	29.185	0.860	0.838	-0.022	2.590
c_{234}	6.3	7.792	1.492	23.678	1.004	0.941	-0.064	6.350
c_{235}	8.1	9.606	1.506	18.590	1.046	1.040	-0.006	0.584
c_{241}	2.1	2.519	0.419	19.932	0.872	1.103	0.232	26.606
c_{242}	6.3	5.796	-0.504	8.002	0.836	1.283	0.447	53.419
c_{243}	10.5	8.988	-1.512	14.400	1.078	1.604	0.526	48.829
c_{244}	14.7	12.161	-2.539	17.273	1.217	1.983	0.766	62.945
c_{245}	18.9	14.637	-4.263	22.556	1.236	2.202	0.966	78.196
c_{351}	1.0	1.030	0.030	2.984	0.183	0.163	-0.020	10.815
c_{352}	3.0	3.071	0.071	2.377	0.210	0.202	-0.008	3.648
c_{353}	5.0	5.099	0.099	1.974	0.278	0.250	-0.029	10.366
c_{354}	7.0	7.129	0.129	1.846	0.305	0.300	-0.004	1.396
c_{355}	9.0	9.148	0.148	1.643	0.355	0.352	-0.003	0.784
c_{361}	2.0	1.896	-0.104	5.223	0.347	0.366	0.019	5.339
c_{362}	6.0	5.931	-0.069	1.158	0.478	0.578	0.100	20.904
c_{363}	10.0	9.854	-0.146	1.459	0.692	0.871	0.179	25.816
c_{364}	14.0	13.783	-0.217	1.551	1.031	1.177	0.146	14.143
c_{365}	18.0	17.669	-0.331	1.841	1.267	1.488	0.221	17.483

Note. |% B| = absolute percent bias; SE_{Pop} is the *SD* of the estimates. Posterior mode estimation results for 30 replicates.

Chapter V

DISCUSSION

5.1. Summary of Findings

This study considered the effects of using examinee-selected items under a comprehensive psychometric model of constructed response item rating. Three hypothetical manners in which examinees may select items when presented with assessments of this type were considered. The recovery of examinee, item, and rater parameters for the HRM-SDT in simulations where examinees selected one of two possible test items and all answered a third were examined and compared with recovery when examinees answered all simulated items. For completeness, an IRT model was also estimated in each condition to serve as a basis for comparison to the estimated proficiency from the HRM-SDT.

The three examinee item selection conditions were meant to span the breadth of possible scenarios, from optimal to the least-favorable in terms of expected model parameter recovery. In the best case—Condition 1—examinees randomly and with equal frequency chose from between the pair of possible items. This met the assumptions of missing completely at random, so model parameters were expected to be recovered with little bias and acceptable variance. Examinee proficiency (θ) was recovered under the HRM-SDT with negligible bias and acceptable RMSE and this held true for examinees choosing either Item 1 or Item 2. All item discrimination parameters (a_i) and nearly 80% of item location parameters (b_{ik}) were recovered with at most moderate bias. Rater parameters were recovered quite well too in this condition, with all rater discrimination parameters (d_{ij}) and over 90% of rater criterion parameters (c_{ijk}) recovered with no more than moderate bias.

The next condition under consideration—Condition 2—was one in which examinees selected items due to a latent trait independent of proficiency called test wisdom. What primarily

distinguished this from Condition 1 was that examinees selected Item 1 (the easier item) about three times as often as Item 2. Just as with Condition 1, there were no discernible differences in the bias or variance of the estimates of examinee proficiency from the HRM-SDT, whether by selected item or the underlying test wisdom indicator. In terms of item parameter recovery, all a_i estimates were recovered with negligible bias and over 85% of the b_{ik} estimates were recovered with at most moderate bias. Rater parameter recovery was slightly poorer than in Condition 1, with 17% and 23% of d_{ij} estimates and c_{ijk} estimates, respectively, showing severe bias.

The final condition—Condition 3—was simulated to represent a worst case scenario, where more proficient examinees made wiser item selections. In such a case, examinees' standing on the very trait to be estimated—subject area proficiency—enabled them to discern item difficulty and hence choose the easier item. This was a clear violation of the assumptions about the missing data mechanisms inherent to most statistical packages' estimation routines, so the parameters were not expected to be recovered as well as they were in either of the previous two conditions. There were differences in the true mean proficiency between examinees selecting Items 1 and 2, so the differential bias and RMSE existed for the proficiency estimates of either group was not surprising. In particular, those selecting Item 2—the harder item—tended to be positively biased, while the proficiency of those making the more astute selection of the easier Item 1 was estimated with a slight negative bias.

Turning to item parameter estimates, all a_i estimates had at least moderate bias and almost 80% of the b_{ik} estimates had at least moderate bias, with the most severe problems being associated with Item 2—the less-frequently selected and harder item. Along those same lines, the signal detection theory rater component of the model suffered poor recovery for the two raters associated with Item 2, but performed adequately for the other four raters. In particular, among the pair of raters rating Item 2, one of the two raters' d_{ij} estimates had severe

bias and nine of the ten raters' criteria (c_{ijk}) estimates had severe bias. On the other hand, the remaining four raters' d_{ij} and 17 of 20 c_{ijk} were recovered with negligible bias.

Implications for Practitioners. In addition to these psychometric findings, there is much to be gleaned with respect to the operational use of examinee-selected items and the practical results of such a practice. Existing research (e.g., DeCarlo et al., 2011; Patz et al., 2002) has shown how rater parameters may be estimated, in addition to item and examinee parameters. The separate rater discrimination and criteria that the HRM-SDT of DeCarlo et al. (2011) enable practitioners to monitor raters' performance much more closely and more meaningfully than the use of aggregate statistics like kappa or inter-rater agreement. This benefit far outweighs the possible additional cost of having a large proportion of items rated twice. Indeed, such rater parameter estimates may enable test publishers to identify gold-standard raters who may be enlisted to train new raters or contribute to the continued development of rubrics. Since this research has confirmed many issues previously raised with using examinee-selected items in the IRT model framework, it is therefore that much more important to ensure that raters perform their tasks well and do not introduce additional error into the measurement process.

These findings have many implications for practitioners interested in using examinee-selected items. The most obvious implication is that items intended for use in an examinee-selected assessment ought to be carefully developed and rigorously pre-tested. In particular, only items whose difficulties are fairly similar ought to appear in a given testlet for examinee selection. In addition, the items should be essentially unidimensional and internally consistent with other test sections.

Practitioners choosing to use examinee-selected items should implement additional analyses as part of an operational testing system. In order to detect whether they are in the unfortunate situation where examinees' proficiency may influence their item selection,

practitioners should scrutinize performance on all sections of the assessment separately for each pattern of selected items. These results should also serve to remind practitioners that when items are selected with greatly varying frequency—e.g., a given item is chosen three times as often as the alternate item—then they may encounter additional issues in model estimation.

The testing conditions also deserve to be considered carefully. The practitioner ought to think carefully about how instructions are written for examinee-selected item conditions, perhaps experimentally investigating a variety of instruction sets and choosing the one that appears most likely to induce students to select the item on which they are likely to perform best. Also, given the relative novelty of this assessment design, perhaps additional time should be given to examinees so that they can read all items under examinee-selection; this might mitigate any test anxiety that may result from the use of this test format.

5.2. Limitations and Direction for Future Research

Despite the great care taken to rigorously answer the research questions addressed in this study, there are some limitations to this work. As a result, a number of additional questions remain to be addressed in future research.

A practical limitation of the study was that it was based on simulated data and a handful of theoretical manners in which examinees may select items. Despite the fact that examinees' item selection mechanisms were carefully chosen, either with support from existing research or to represent best- or worst-case scenarios, a better understanding of examinees' selection behavior is needed to ensure that models such as the HRM-SDT adequately manage the challenges of examinee-selected items.

This study evaluated how well the HRM-SDT and a competing IRT model performed in the presence of examinee-selected items. Specifically, existing models were applied to data

whose missingness was known to violate assumptions implicit to the estimation techniques. A natural next step in this research thread is to extend the HRM-SDT to explicitly include the missing data mechanism. It is hoped that in making that extension, better recovery of examinee, item, and rater parameters may be possible, even in the most extreme case where examinee item selection is closely related to underlying proficiency.

Upon the successful development of such an extended model, experimental research is needed to better understand how examinees behave in the relatively novel setting of examinee-selected items. Do students choose rationally and attempt to maximize their expected test score? Are they more likely to choose the first item presented for selection? This study, along with the areas proposed for future work, may yield sufficient evidence in support of the use of examinee-selected items and ultimately a more widespread use of this assessment design.

REFERENCES

Allen, N. L., Holland, P. W., & Thayer, D. T. (2005). Measuring the benefits of examinee-selected questions. *Journal of Educational Measurement*, *42*(1), 27–51. doi:10.1111/j.0022-0655.2005.00003.x

Bell, J. (1997). Question choice in English literature examinations. *Oxford Review of Education*, *23*(4), 447–458. doi:10.1080/0305498970230402

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores*. Oxford, England: Addison-Wesley.

Bock, D. R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. doi:10.1007/BF02291411

Bradlow, E. T. & Thomas, N. (1998). Item response theory models applied to data allowing examinee choice. *Journal of Educational and Behavioral Statistics*, *23*(3), 236–243. doi:10.3102/10769986023003236

Bridgeman, B., Morgan, R., & Wang, M. M. (1997). Choice among essay topics: Impact on performance and validity. *Journal of Educational Measurement*, *34*(3), 273–286. doi:10.1111/j.1745-3984.1997.tb00519.x

Campbell, J. R. & Donahue, P. L. (1997). Students selecting stories: The effects of choice in reading assessment. *National Center for Education Statistics, 97-491*. Washington, DC.

Chi, M. T. H. (1978). Knowledge structures and memory development. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 73–96). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance (Cambridge Handbooks in Psychology)* (1st ed., pp. 21–30). New York: Cambridge University Press.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 7–70). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cicchetti, D. & Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, *11*(3), 101–109.

College Board. (2006). *Test analysis: Advanced Placement chemistry*. New York.

College Board. (2010a). *Test analysis: Advanced Placement United States history*. New York.

College Board. (2010b). *Test analysis: Advanced Placement European history*. New York.

DeCarlo, L. T. (2002). A latent class extension of signal detection theory, with applications. *Multivariate Behavioral Research*, *37*(4), 423–451. doi:10.1207/S15327906MBR3704_01

DeCarlo, L. T. (2008). *Studies of a latent-class signal-detection model for constructed-response scoring (ETS Research Report RR-08-63)*. Princeton, NJ: Educational Testing Service. Retrieved from: http://www.ets.org/Media/Research/pdf/RR-10-08.pdf.

DeCarlo, L. T., Kim, Y. K., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, *48*(3), 333–356. doi:10.1111/j.1745-3984.2011.00143.x

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mawah, NJ: Lawrence Erlbaum Associates, Inc.

Fitzpatrick, A. R. & Yen, W. M. (1995). The psychometric characteristics of choice items. *Journal of Educational Measurement*, *32*(3), 243–259. doi:10.1111/j.1745-3984.1995.tb00465.x

Flora, D. B. & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–91. doi:10.1037/1082-989X.9.4.466

Gabrielson, S., Gordon, B., & Engelhard Jr., G. (1995). The effects of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education*, *8*(4), 273–290. doi:10.1207/s15324818ame0804_1

Galindo Garre, F. & Vermunt, J. K. (2006). Avoiding boundary estimates in latent class analysis by bayesian posterior mode estimation. *Behaviormetrika*, *33*(1), 43–59. doi:10.2333/bhmk.33.43

Gee, T. W. (1987). Drafting and revising processes in grade 12 students ' examination writing. *Alberta Journal of Educational Research*, *33*(2), 96–114.

Hamp-Lyons, L. & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, *3*(1), 49–68. doi:10.1016/1060-3743(94)90005-1

Holman, R. & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1–17. doi:10.1111/j.2044-8317.2005.tb00312.x

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers' choice: an investigation of the effect of topic on language-test performance. *Language Testing*, *16*(4), 426–456. doi:10.1177/026553229901600402

Kim, Y. K. (2009). *Combining constructed response items and multiple choice items using a hierarchical rater model* (Doctoral dissertation). Teachers College, Columbia University, New York.

Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. doi:10.1037/0022-3514.77.6.1121

Linn, R. L., Betebenner, D. W., & Wheeler, K. S. (1998). *Problem choice by test takers: Implications for comparability and construct validity (Center for the Study of Evaluation Technical Report 485)* (Vol. 1522). Los Angeles, CA. Retrieved from: http://cse.ucla.edu/products/reports/TECH485.pdf.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (Second Ed.). New York: Wiley-Interscience.

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, *48*(3), 477–482. doi:10.1007/BF02293689

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, *31*(3), 234–250. doi:10.1111/j.1745-3984.1994.tb00445.x

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, *25*(3), 707–726. doi:10.1177/001316446502500304

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. doi:10.1177/014662169201600206

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Myford, C. M. & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

OCR. (2003). *General Certificate of Secondary Education (GCSE) in English Literature, Second Edition*. *English*. Cambridge, United Kingdom.

Patz, R. J. (1996). *Markov chain Monte Carlo methods for item response theory models with applications for NAEP* (Doctoral dissertation). Carnegie Mellon University, Pittsburgh, PA.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341–384. doi:10.3102/10769986027004341

Powers, D. E. & Bennett, R. E. (1999). Effects of allowing examinees to select questions on a test of divergent thinking. *Applied Measurement in Education*, *12*(3), 257–279. doi:10.1207/S15324818AME1203_3

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer New York. doi:10.1007/978-0-387-89976-3

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. doi:10.2307/2335739

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17)*. Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research*, *49*(2), 252–279. doi:10.3102/00346543049002252

Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577. doi:10.1007/BF02295596

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science (New York, N.Y.)*, *185*(4157), 1124–31. doi:10.1126/science.185.4157.1124

Vermunt, J. K. & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced*. Belmont, MA: Statistical Innovations, Inc. Retrieved from http://www.statisticalinnovations.com/products/LGtechnical.pdf

Vermunt, J. K. & Magidson, J. (2008). *LG-Syntax user's guide: Manual for Latent GOLD 4.5 syntax module*. Belmont, MA: Statistical Innovations Inc. Retrieved from http://www.statisticalinnovations.com/technicalsupport/LGSyntaxusersguide.pdf

Wainer, H. & Thissen, D. (1994). On examinee choice in educational testing. *Review of Educational Research*, *64*(1), 159–195. doi:10.3102/00346543064001159

Wainer, H., Wang, X. B., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinees choice? *Journal of Educational Measurement*, *31*(3), 183–199. doi:10.1111/j.1745-3984.1994.tb00442.x

Wang, W. C., Jin, K. Y., Qiu, X. L., & Wang, L. (2012). Item response models for examinee-selected items. *Journal of Educational Measurement*, *49*(4), 419–445. doi:10.1111/j.1745-3984.2012.00184.x

Wang, X. B., Wainer, H., & Thissen, D. (1995). On the viability of some untestable assumptions in equating exams that allow examinee choice. *Applied Measurement in Education*, *8*(3), 211–225. doi:10.1207/s15324818ame0803_2

Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.

Willmott, A. S. & Hall, C. G. W. (1975). *O-level examined: The effect of question choice*. London: Schools Council.